VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ HOMOLOGNÍCH GENŮ POMOCÍ METOD ZPRACOVÁNÍ SIGNÁLŮ
HOMOLOGY SEARCH USING DIGITAL SIGNAL PROCESSING METHODS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Yana Kamar**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Denisa Maděránková, Ph.D.**
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Yana Kamar ***ID:*** 174504
Ročník: 3 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Vyhledávání homologních genů pomocí metod zpracování signálů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na témata: metody vyhledávání homologních genů v prokaryotních i eukaryotních genomech, metody signálové reprezentace nukleotidových sekvencí a metody rychlého porovnávání signálů. 2) V libovolném programovém prostředí implementujte alespoň dvě metody signálové reprezentace sekvencí. 3) Navrhněte způsob porovnávání sekvenčních signálů za účelem vyhledávání homologií. 4) Navrženou metodiku naprogramujte a otestujte na rozsáhlém souboru anotovaných prokaryotních genomů. 5) Výsledky diskutujte a vyhodnoťte úspěšnost vyhledání homologií.

DOPORUČENÁ LITERATURA:

[1] CUI, X.; VINAŘ, T.; BREJOVÁ, B.; SHASHA, D.; LI, M. Homology search for genes. Bioinformatics. 2007, 23(13), i97-i103.

[2] PEARSON, W. An introduction to sequence similarity ("homology") searching. Curr. Protoc. Bioinformatics. 2013, Chapter 3, Unit3.1.

Termín zadání: 6. 2. 2017 ***Termín odevzdání:*** 02.06.2017

Vedoucí práce: Ing. Denisa Maděránková, Ph.D.
Konzultant:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ Fakulta elektrotechniky a komunikačních technologií
Ústav biomedicínského inženýrství
-2-

UPOZORNĚNÍ:

Autor bakalářské práce práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb

ABSTRAKT

Práce obsahuje teoretický úvod do molekulární biologie a genetiky na potřebné úrovní, včetně popisu stavby DNA a homologních genů. Jsou popsané pevné a fyzikálně-chemické typy mapování nukleotidů, metody zpracování digitálních signálů. V prostředí MATLAB byli naprogramované numerické reprezentace genů, jmenovitě: rozbalená a kumulovaná fáze, denzitní vektory. S použitím rozbalené fáze a denzitních vektorů s různou délkou bylo uskutečněno vyhledávání CDS úseku v celém genomu pomocí metrických vzdálenosti (euklidovské a canberrské) a korelace. Dále s použitím metrických vzdálenosti byl vyhledán homologní gen v více či méně podobných bakteriálních genomech. Výsledkem je orientační práh vzdálenosti(euklidovské a canberrské) pro nalezení homologních genů v genomech.

KLÍČOVÁ SLOVA

DNA, RNA, homologní gen, numerická reprezentace, numerická mapa, BLAST, euklidovská vzdálenost, canberrská vzdálenost, digitální zpracování signálů

ABSTRACT

Thesis includes the theoretical introduction to molecular biology and genetics on the necessary level, including a description of the structure of DNA and the homologous gene. Described are fixed and physic-chemical kinds of nucleotide mapping, methods for processing digital signals. Numerical representations of genes that were programmed in MATLAB: unwrapped and accumulated phases, density vectors. Using the unwrapped phase and density vectors with windows of different lengths was performed CDS searching in the entire genome by calculation metric distances (euclidean and canberian) and correlation. Also, using the metric distances, a homologous gene was found in more or less similar bacterial genomes. The result is the approximate threshold of distance (euclidean and canberian) using to find homologous genes in genome.

KEYWORDS

DNA, RNA, homologous gene, numerical representation, numerical map, BLAST, euclidean distance, canberian distance, digital signal processing

`

KAMAR, Y. *Vyhledávání homologních genů pomocí metod zpracování signálů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2017. 57 s. Vedoucí práce Ing. Denisa Maděránková, Ph.D.

`

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Vyhledávání homologních genů pomocí metod zpracování signálů jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucí bakalářské práce Ing. Denisa Maděránková, Ph.D. za rady a vstřícnost při konzultacích.

V Brně dne ....

(podpis autora)

`

OBSAH

`

`

ÚVOD

Tato bakalářská práce se zabývá vyhledáváním homologních genů pomocí metod zpracování signálů.

V dnešní době databázi obsahují obrovské množství dát a s časem se toto množství pouze roste. Některé geny jsou důkladně prostudované. Jejích funkce je známa. V případě nalezení nového genu se nemůžeme ručit, jak je tento gen v organismu využíván. Právě homologní geny jsou klíčem k porozumění funkcí neznámých genů. Jsou to totiž příbuzné geny, které lze najít porovnáním se známými geny v databázi.

Před vyhledáním pomocí metod zpracovaní signálů je potřeba sekvenci nukleotidů převést vhodnou metodou na číslicový signál. Takový převod avšak nesmí zavést chybnou informaci. Výsledný signál pak musí být pro danou sekvenci jedinečný a ideálně by měl zohledňovat vlastnosti nukleotidů.

Jako vhodné numerické reprezentace byli vybrání rozbalená fáze, kumulovaná fáze a také denzitní vektory.

Jako vhodné metody vyhledání homologních genů byli vybrání metrické vzdálenosti(euklidovská a canberrská) a korelace. Po testovaní se ukázalo, že kumulovaná fáze nesplňuje vše požadavky pro vhodnou numerickou reprezentaci a korelace pak nesplňuje požadavky pro vyhledání samotné.

Výsledkem bakalářské práce je program, který problematiku vyhledání homologních genů na určité úrovní řeší. Dalším výsledkem je orientační práh užitečné vzdálenosti(canberrské a euklidovské).

`

1 ZÁKLADY MOLEKULÁRNÍ BIOLOGIE A GENETIKY

1.1 Prokaryotické a eukaryotické buňky

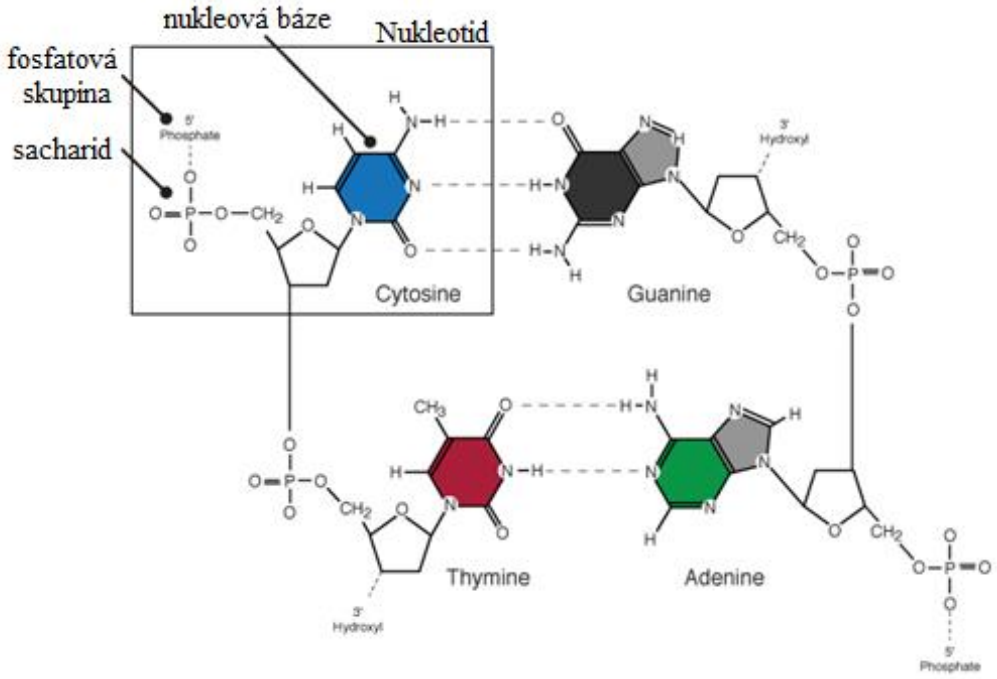
Každý živý organizmus, kromě virů, se skládá z buněk. Rozlišují se však mezi sebou množstvím a druhém buněk, ze kterých jsou sestavené. Pokud se jedná o druzích, existuje dvě základní skupiny: eukaryotické a prokaryotické buňky.

Prokaryotické jsou mnohem starší, mají je především menší organizmy, jako jsou bakterie, archea, sinice. Kruhová DNA u prokaryotů není izolována membránou, nachází se přímo v cytoplasmě. Na rozdíl od něj, eukaryota mají jádro a další membránové organely, kromě toho jsou větší. Eukaryotické organizmy se často skládají z většího počtů buněk než prokaryotické.

1.2 Nukleové kyseliny

U většiny organizmů, jak prokaryotických, tak i eukaryotických, genetická informace je uložená v makromolekulách DNA. Menšinu pak tvoří skupina prokaryot a virů, u kterých touž funkci plní molekula RNA.

DNA, stejně jako RNA, je nukleová kyselina, skládající se z podjednotek zvaných nukleotidy. Každý nukleotid se pak skládá ze následujících časti: nukleová báze, fosfátová skupina a pětiuhlíkatý monosacharid (deoxyribóza v případě DNA a ribóza v případě RNA).(viz Obr. 1.1)

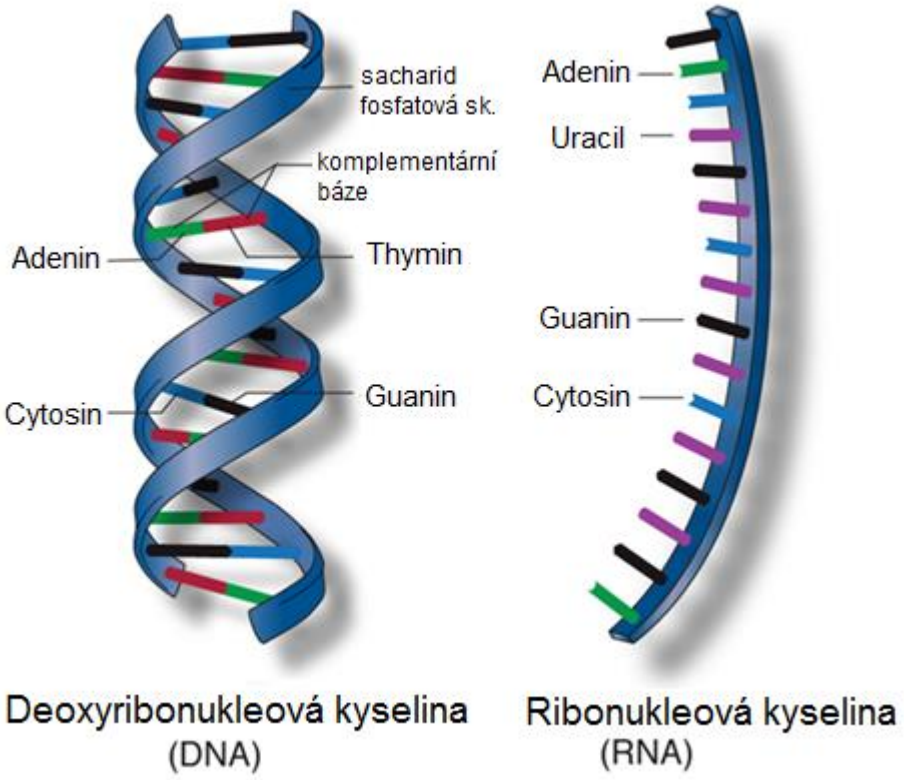


Obr. 1.1 – Nukleotidy

`

Dusíkaté báze se rozděluji na dvě skupiny: purinové (obsahuji bicyklické báze: adenin a guanin) a pyrimidinové (obsahuji monocyklické báze: cytozin, tymin, uracil). Tymin lze nalézt pouze v molekulách DNA. V RNA tuto bázi nahrazuje uracil.

Popsané dříve podjednotky nukleotidů jsou spojené v dlouhé řetězy, které mají šroubovitý tvar. DNA obsahuje dva komplementární vlákna, RNA pouze jedno. Na základě komplementarity se řetězy DNA spojují přes nukleové báze. Adenin je komplementární s tyminem (v případě DNA) a uracilem (v případě RNA). Tvoří dvě vodíkové vazby. Guanin je komplementární s cytozinem a tvoří tří vodíkové vazby. (viz Obr. 1.2)



Obr. 1.2 – Struktura DNA a RNA

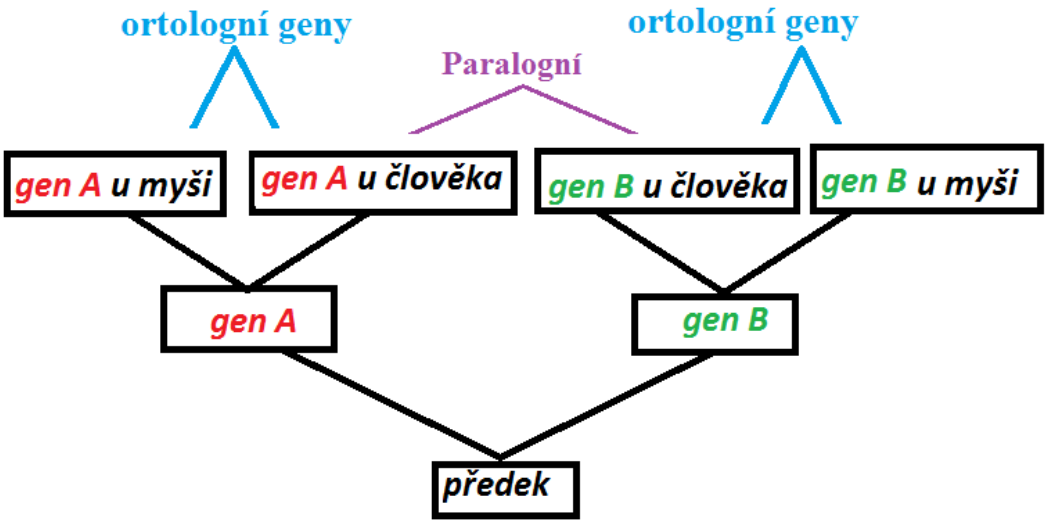
1.3 Homologní geny

Gen je úsek DNA kódující bílkovinu nebo RNA. Každý gen má exony (části, které se replikuji a podílejí se na tvorbě bílkoviny nebo RNA) a introny (zbytek genu, který spojuje exony). Úkol intronů není dosud zcela jistě znám.

Pojmem homologní geny se označuji geny odvozené z jednoho společného genu. Rozlišují se dva typy homologních genů.

První typ tvoří ortologní geny – geny u různých druhů, které se vyvinuly ze společného předka. Paralogní geny jsou výsledkem duplikace genu předka. Orthology zachovávají stejnou funkci v průběhu evoluce, zatímco paralogy vyvinuli nové funkce. (viz Obr. 1.3)

`



Obr. 1.3 – Ortologní a paralogní geny schematicky.

Vzhledem k tomu, že společný gen předka často není k dispozici, nelze posoudit, zda jeden gen je opravdu homologní s druhým. Avšak když dvě sekvence jsou si hodně podobní, lze předpokládat, že tyto sekvence nevznikli nezávisle na sobě, ale mají společného předka. Tato skutečnost se používá k vyhledávání homologních genů v databázích.

Důvodů, proč vyhledávání homologů je důležitým a diskutovaným problémem, je několik. Zaprvé v dnešní době věda pokročila značně dopředu. S časem se vynalézají nové proteiny a geny. Interpretace nukleotidových posloupnosti a proteinů, to znamená zjištění jejích účelů v organizmu, není zcela jednoduchá.

Jelikož příbuzné geny a proteiny mohou mít stejnou funkci, při vynálezu nového genu prvním krokem bude vyhledávání homologních genů v databázích už známých posloupnosti nukleotidů(nebo proteinů). Při tom nastává další problém: objemy dat v takových databázích je obrovský. Nástroj pro vyhledaní měl by být výpočetně jednoduchý, rychlý a zároveň robustní a dostatečně výkonný. Další kapitoly jsou věnované způsobům a nástrojům, které se používají pro účel vyhledávaní homologních genů, a to nejen metodou zpracovaní signálů.

`

2 NUMERICKÉ REPREZENTACE NUKLEOVÝCH KYSELIN

Jelikož tato práce se zabývá vyhledáváním homologů pomocí metody zpracování signálů, před zpracováním samotným je potřeba převést posloupnost DNA (nebo RNA) na číslicový signál pomocí vhodného typu mapování.

Většina zdrojů rozděluje mapování na 2 typy: pevné mapování (angl. Fixed mapping) a mapování na základě fyzikálně-chemických vlastností (angl. Physico Chemical Property Based Mapping). V dané bakalářské práci se toto rozdělení taky bude dodržováno.

2.1 Pevné mapování

U metod pevného mapování nukleotidová posloupnost se transformuje na libovolnou číselnou posloupnost. Tuto skupinu metod tvoří následující reprezentaci:

- Voss [1]
- Reprezentace čtyřstěnem [2]
- Celočíselná [3]
- Pomocí reálných čísel [4]
- Pomocí komplexních čísel [5]
- Pomocí denzitních vektorů [6]
- 1. a 4. kvadrant [7]
- Rozbalena a kumulovaná fáze [3]

Tab. 2.1 – Pevné mapování. Ukázka

Bod	*Reprezentace*	*S(n)=[CGAT]*
A	*Voss*	u_C *=[1, 0, 0, 0]* u_G *=[0, 1, 0, 0]* u_A *=[0, 0, 1, 0]* u_T *=[0, 0, 0, 1]*
B	*Celočíselná*	*x[n]=[1,3,2,0] nebo x[n]=[2,4,3,1]*
C	*Pomocí reálných čísel*	*x[n]=[0.5, -0.5, -1.5, 1.5]*
D	*Pomocí komplexních čísel*	*x[n]=[-1+j, -1-j, 1+j, 1-j]*
E	*1. a 4. kvadrant*	$u_x = \left[\frac{\sqrt{3}}{2}, \sqrt{3}, \sqrt{3}+\frac{1}{2}, \sqrt{3}+1\right]$ $u_y = \left[\frac{1}{2}, 0, -\frac{\sqrt{3}}{2}, 0\right]$

`

Voss

Voss je jednou z nejjednodušších metod. Posloupnost převádíme na 4 indikační vektory u_T, u_A, u_C, u_G. Pro jeden konkrétní nukleotid v posloupnosti platí, že v odpovídajícím jemu vektoru na jeho pozici je 1, když v ostatních vektorech na této pozici budou 0 (viz tab. 2.1 bod A). Daná metoda nezohledňuje vlastnosti nukleotidů, pouze poskytuje informaci o jejích poloze a četnosti.

Reprezentace čtyřstěnem

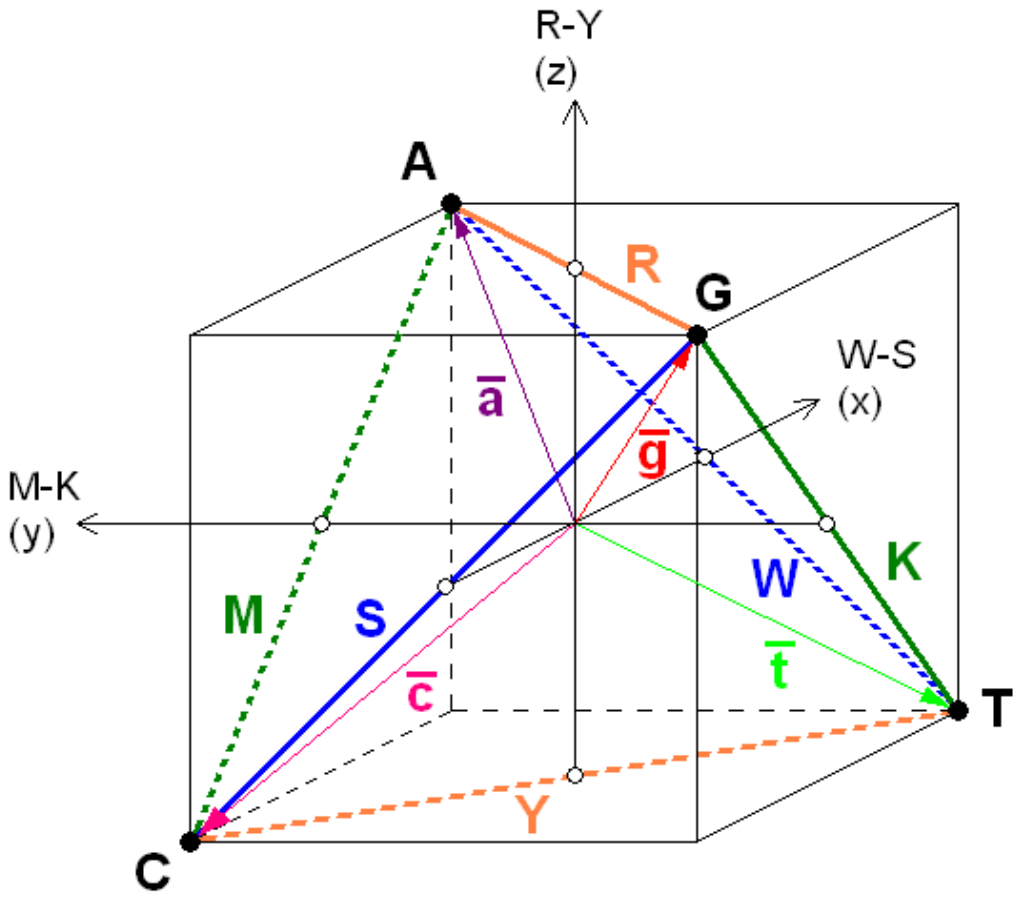
Mapování pomocí čtyřstěnu znázorňuje Obr. 2.1. Nukleotidy jsou umístěné v 3D prostoru v rozích čtyřstěnu. Prostor se popisuje pomocí os x-y-z, které se protínají ve středu čtyřstěnu. Každá z 6 projekci obsahuje informaci o jedné biochemické vlastnosti: silná-slabá vazba(S/W), amin-keto skupina(M/K), purin-pirimidin(R/Y). Přičemž každý pár vlastnosti se nachází na kolmých projekcích. Pokud vzdálenost od počátku k rohům se rovná 1, platí, že souřadnice nukleotidů jsou následující:

A = (1, 1, 1)

C = (-1, 1, -1)

G = (-1, -1, 1)

T = (1, -1, -1)



Obr. 2.1. – Nukleotidový čtyřstěn [6]

`

Celočíselná reprezentace

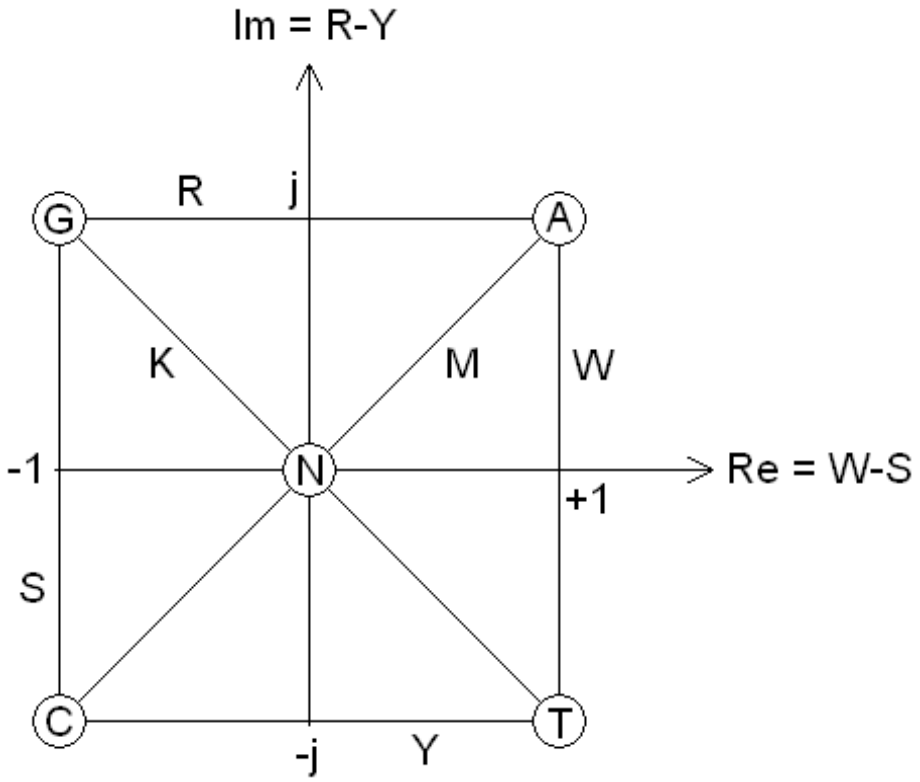
Mapování pomocí celých čísel nahrazuje posloupnost nukleotidů posloupnosti čísel, přičemž nukleotidy budou nahrazené následně : A=1, C=2, T=3, G=4 nebo A=0, C=1, T=2, G=3. Tato metoda stejně jako Voss sděluje pouze informaci o poloze a četnosti určitých nukleotidů. Kromě toho ale zavadí i chybnou informaci, že A<C<T<G, proto je nevhodná pro analýzy, jen pro předzpracování. (tab. 2.1 bod B).

Reprezentace reálnými čísly

Podobnou metodou je mapování pomocí reálných čísel: A = -1,5, T = 1,5, C = 0,5 a G = -0,5. Daná metoda nese v sobě informaci o komplementaritě bází, protože absolutní hodnoty komplementárních bází jsou stejné, kromě toho pyrimidinové báze jsou kladné, když purinové naopak záporné. (tab. 2.1 bod C).

Reprezentace komplexními čísly

Metoda s využitím komplexních čísel také odráží komplementaritu A-T a C-G. Nukleotidy se nahrazují komplexními čísly: A = 1+j, C = 1+j, G = -1-j, T = 1-j, což lze grafický zobrazit (viz Obr. 2.1). Osa x je reálnou složkou, osa y je imaginární složkou. Komplementarita je vyjádřená symetrii podle osy x(Re). Purinové a pyrimidinové báze jsou umístěné symetricky podle osy y(Im). (tab. 2.1 bod D).



Obr 2.2 - Reprezentace pomocí komplexních čísel [6]

`

Reprezentace denzitními vektory

Výpočet denzitních vektorů je další typ pevného mapování. Vychází z metody Voss. Prvním krokem se vytvářejí indikační vektory u_T, u_A, u_C, u_G. Dělá se to tak, že na konkrétní pozici je 1 v tom indikačním vektoru, který odpovídá příslušnému nukleotidů. V ostatních vektorech na této pozici jsou 0. Na rozdíl od Voss mapování lze tady použít i nepřesně definované nukleotidy, které jsou uvedené v tab. 2.2 :

Tab. 2.2. - Nepřesně definované nukleotidy

Zkratka (vlastnost)	*Nukleotidy*
S (Silná vazba)	*C,G*
W (Slabá vazba)	*A,T*
R (Puriny)	*A,G*
Y (Pirimidin)	*C,T*
M (Aminoskupina)	*A,C*
K (Ketoskupina)	*G,T*
B	*C,G,T*
D	*A,G,T*
H	*A,C,T*
N	*A,C,T,G*

Dělá se to tak: hodnota 1 (neboli pravděpodobnost výskytu nukleotidu 100%) se rozdělí na množství možných nukleotidů a získané číslo se dává do každého odpovídajícího příslušným nukleotidům indikačního vektoru. V zbývajících vektorech na této pozici je 0. Např. pro symbol K, který má stejnou pravděpodobnost výskytu T a G: u_T =0,5, u_C =0, u_G =0,5, u_A =0. Pro symbol N: u_T =0,25, u_C =0,25, u_G =0,25, u_A =0,25.

Dalším krokem je průměrování každého indikačního vektoru s posuvným oknem určité délky W. Posouvá se vždy o jeden nukleotid. Ve výsledku denzitní vektor je kratší než původní sekvence o (W-1) hodnot, proto je potřeba tuto chybu eliminovat. Jedním ze způsobu jak zachovat délku vektoru je přidaní dodatečných hodnot k indikačním vektorem před průměrováním samotným. Hodnoty se přidávají na začátek a na konec vektorů, proto by bylo vhodnější použit lichou délku okna, aby následně (W-1) bylo sudé a délka přidaných hodnot se na začátku a na konci vektorů rozdělila rovnoměrně. Přidávat se bude hodnota 0,25 a to do každého indikačního vektorů ze čtyř. Bude to symbolizovat stejnou pravděpodobnost výskytu jakéhokoliv nukleotidu na pozicích před a po analyzované sekvenci. Výsledný vzorec pro vypočet denzitních vektorů vypadá takto:

$$d_x[n] = \frac{\sum_{i=n-W/2}^{n+W/2} u_x[i]}{W} \qquad (1)$$

Kde W je délka okna, n je pozice v denzitním vektoru a X je jeden ze 4 nukleotidu.

Denzitní vektory neposkytují přesnou informaci o pozici nukleotidů v řetězci DNA (je přítomná ztráta informace), ale zobrazuje ho jako celek, protože aktuální hodnota v DV je závislá na sousedních hodnotách. Změna délky okna mění míru, z jakou vzdálenější nukleotidy

`

působí na aktuální hodnotu. Výběr okna může se měnit v závislosti na metodě zpracování signálů, která se použije pro analýzu. Tato metoda by mohla být použitá pro vyhledávání homologních genů, protože vytváří téměř jedinečný obrazec pro danou sekvenci. Kód této metody je v příloze Kódy (bod 1).

1. a 4. kvadrant

Metoda 1. a 4. kvadrantu je grafickou metodou, pomocí které se dá znázornit průběh sekvence DNA nebo RNA v 2D prostoru. Tento prostor se popisuje pomocí os x-y. Každý nukleotid je nahrazen vektorem, který se začíná v bodě (0,0) a končí v 1. nebo 4. kvadrantu. (Viz Obr. 2.5) Souřadnice pro různé typy nukleotidů jsou následující:

$$A = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) C = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right) G = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) T = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$$

Aby se to mohlo zobrazit jako křivka na grafu, musí být každá následující hodnota získaná přičítáním předcházející k hodnotě příslušného nukleotidu. Díky použiti výhradně 1. a 4. kvadrantu se netvoří smyčky a reprezentace tím pádem není degenerativní. Kromě toho purinové báze mají kladnou hodnotu na y-ové ose a pyrimidinové naopak zápornou. Pokud v nějakém úseku bude převažovat jeden druh nukleotidů a v dalším druhý, pak to bude patrné na grafu (jako vzrůstající a klesající částí).

Reprezentace s využitím fáze

Rozbalena a kumulovaná fáze je další možnost přeložení posloupnosti nukleotidů na číslicový signál. Vychází z metody s použitím komplexních čísel, protože výsledná sekvence v komplexních číslech má fázovou charakteristiku. Pro komplexní hodnoty A,C,G,T={ 1+j, −1−j, −1+j, 1−j} fáze nukleotidů mohou nabývat pouze hodnot {−3π/4, −π/4, π/4, 3π/4} radián. Kumulovanou fázi se nazývá vektor, ve kterém hodnoty se počítají podle vzorce [4]:

$$\theta_c = \frac{\pi}{4}[3(n_G - n_C) + (n_A - n_T)] \quad (2)$$

Kde n_C, n_G, n_A *a* n_T je počet cytozinu, guaninu, adeninu a tyminu od první do aktuální pozici.

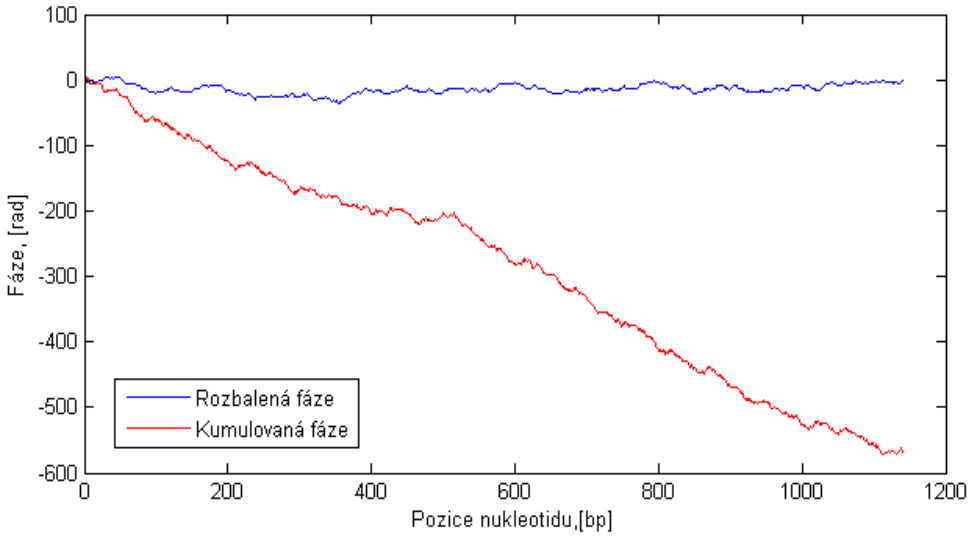
Rozbalená fáze je upravená absolutní hodnota rozdílu mezi aktuální hodnotou a předcházející. Upravená je přidáním nebo odečtením odpovídajícího násobku 2π k nebo od fáze aktuálního prvku tak, aby výsledné hodnoty byli menší než π. Upraví se provádí následujícím způsobem[3]:

$$\theta_u = \frac{\pi}{2}[(n_+ - n_-)] \quad (3)$$

Kde n_+ je pozitivní změna, n_- je negativní přechod. Negativnimi jsou přechody A → T, T → C, C → G, G → A. Pozitivní pak jsou A → G, G → C, C → T, T → A.

Kód této metody je v příloze Kódy (bod 2 a 3). Grafická ukázka kumulované a rozbalené fáze genu ukázaná na Obr. 2.3.

`



Obr. 2.3 – Rozbalená a kumulovaná fáze mitochondriálního genu *mus musculus* (viz 4. kapitolu)

Rozbalenou fázi lze také vytvořit následujícím způsobem: nukleotidy se nahrazují komplexními čísly: A = 1+j, C = 1+j, G = -1-j, T = 1-j. Pak komplexní čísla se nahrazují uhlem, který nukleotidy svírají (viz. Obr. 2.2).

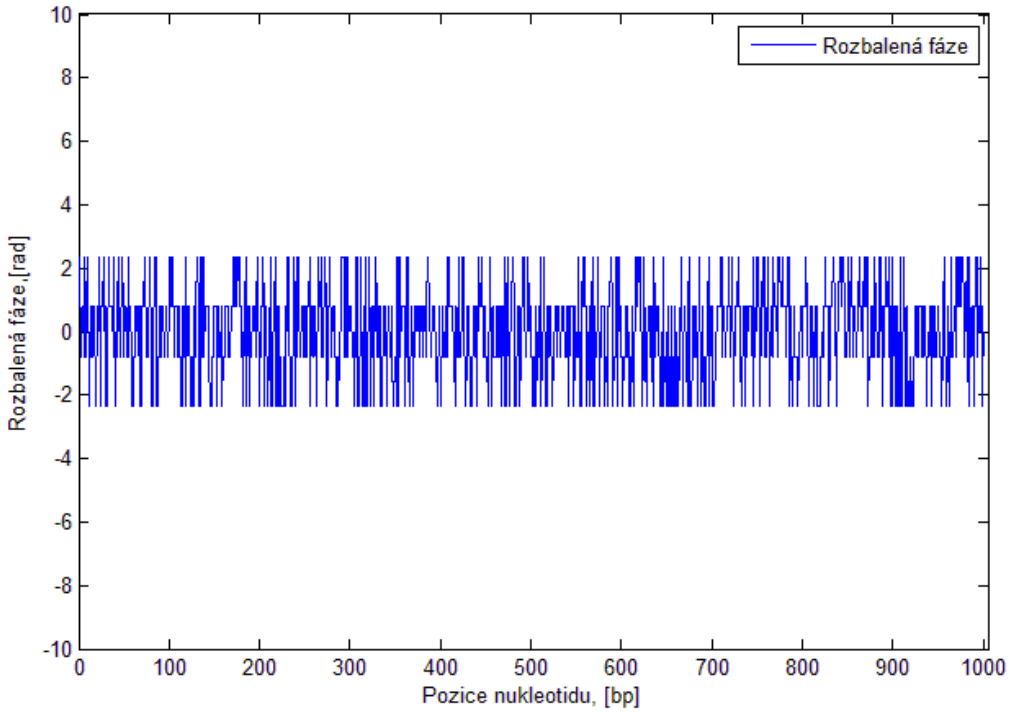
Jelikož komplementární nukleotidy jsou naproti sebe, lze signál pro komplementární vlákno odvodit od signálu pro kódující podle vztahu:

$$RF(komp.) = RF.(-1) \tag{4}$$

Třeba také poznamenat, že výsledek je ve špatném směru. Celý vektor je nutno otočit.

Příklad signálu délkou 1000 nukleotidu je na Obr. 2.4:

`



Obr. 2.4 – Rozbalená fáze prvních 1000 nukleotidů mitochondriálního genu *mus musculus* (viz 4. kapitolu)

2.2 Mapování na základě fyzikálně-chemických vlastností

U těchto metod mapování se biofyzikální a biochemické vlastnosti DNA používají pro mapování DNA sekvence. Mapování na základě fyzikálně-chemických vlastností zahrnuje metody:

- EIIP [10]
- Pomocí atomového čísla [11]
- Pomocí spárovaných nukleotidů [9]
- DNA-walking [12]
- Reprezentace Z-křivkou [13]

EIIP a atomové číslo

Energie delokalizovaných elektronů v nukleotidech se nazývá angl. electron-ion interaction potential (EIIP). V metodě s použitím EIIP se nukleotidy nahrazují odpovídajícími hodnotami energie: A = 0.1260, C =0.1340, G = 0.0806, a T = 0.1335. Analogicky se dá nahradit posloupnost nukleotidů atomovými čísly následujícím způsobem: A = 70, C = 58, G = 78 a T = 66.

Reprezentace spárovanými nukleotidy

Metoda spárovaných nukleotidů spojuje páry A-T a C-G tím, že první pár se nahrazuje hodnotou +1, druhy pár hodnotou -1. Pokud se to pouze nahrazuje, vzniká jeden indikační vektor. Existuje také možnost vytvořit 2 indikační vektory pro každý pár zvlášť.

DNA-walking

Existuje metoda podobná předcházející metodě, která se nazývá DNA-walking. Nejčastěji se to používá jako grafická metoda pro znázornění průběhu DNA. Graf pokračuje směrem nahoru (přidává se hodnota +1) v případě, že nukleotid je pyrimidinový (C-T) nebo směrem dolů (přidává se hodnota -1), je-li nukleotid purinový (A-G). Tyto reprezentace redukují informační obsah sekvence.

Reprezentace Z-křivkou

Další numerická reprezentace je reprezentace z-křivkou. Z-křivka je 3-D křivka, která nabízí jedinečnou reprezentaci pro vizualizaci a analýzu DNA sekvence. Tři složky Z-křivky, {xn, yn, zn}, reprezentují tři nezávislé nukleotidové rozdělení, které zcela popisují DNA sekvenci. Složky xn, yn, zn ukazuje distribucí purinových bázi oproti pyrimidinovým (R proti Y), amino proti keto skupině (M proti K), a silné H-vazby proti slabým W-vazbam (S proti W).

`

3 VYHLEDÁVANÍ HOMOLOGNÍCH GENŮ

Homologní geny, jak už bylo zmíněno dříve, jsou příbuzné geny, které mají společného předka. To znamená, že posloupnosti jsou velice podobné. Pak pro vyhledávání homologních genů můžou být použité metody zarovnání nebo počítání míry podobnosti.

Před popisováním metod zpracování digitálních signálů, kterým je věnovaná bakalářská práce, třeba zmínit i obecné metody běžně používané k tomuto úkolů.

3.1 Obecné metody vyhledávaní (zarovnání) posloupnosti nukleotidů

Široce používanými algoritmy jsou:

- BLAST [13] [14]
- FASTA [15]
- CLUSTAL [16]

BLAST (angl. Basic Local Alignment Search Tool) je široce používány nástroj pro vyhledávaní proteinových a DNA sekvencí v databázích. Je to heuristická metoda, což znamená, že výsledkem je odhad správné odpovědi. Zarovnání jedné sekvence se provádí s více sekvencemi v databázích.

Algoritmus má několik kroku. V prvním kroku sekvence bude rozdělena na úseky určité délky. Ostatní sekvence v databází také budou rozdělená a uložená. Následujícím krokem je vyhledávaní úseků z dané sekvence. Jakmile se objeví nějaká shoda s úsekem v databází, prodlužuje se tento úsek do obou stran a porovnává se z stejně prodluženou sekvenci. Následně se musí provést skórování prodluženého úseku (tzv. HSP). Jestli je skóre postačující, prodlužování pokračuje a skórovaní se provádí zase. Každá zarovnána sekvence na konci také dostává tzv. e-hodnotu (další způsob hodnocení), aby se několik výsledků mohli porovnat mezi sebou.

Programy FASTA hledají oblastí lokální nebo globální podobnosti mezi proteiny nebo DNA sekvencemi, a to buď vyhledáváním proteinů nebo DNA v databází nebo tím, že určí místní zdvojení v sekvenci. Stejně jako BLAST, pomocí FASTA lze odvodit funkční a evoluční vztah mezi sekvencemi.

CLUSTAL je modernější algoritmus pro zarovnání více sekvencí. Principem je zarovnání dvojic sekvencí a přiřazení nějakého skóre podobnosti mezi jednotlivými dvojicemi sekvencí. Nejvíce podobné sekvence se spojují do jedné skupiny a tak se pokračuje dokud nevznikne tzv. guide tree. V tomto stromě se příbuzné sekvence (homologní) mohou nacházet blízko k sobě.

Další programy, které se mohou podílet na vyhledávaní homologů, které avšak nebudou popsané podrobně v této práci: HMMER3 (Hidden Markov Model rychlého heuristického a iteračního HMM postupu vyhledávání), SSEARCH (identifikace společných molekulárních podsekvencí.), MAFFT (nový způsob pro rychlé vícenásobné porovnání sekvencí na základě rychlé Fourierovy transformace), MUSCLE (vícenásobné porovnání sekvencí s vysokou přesností). [18]

`

3.2 Metody zpracovaní signálů

Korelace

První metodou, která bude popsaná v této kapitole a také naprogramovaná, je korelace [19]. Korelační posloupnost je funkce posunutí (teta) jedné posloupnosti vůči druhé, co odráží i vzorec:

$$r_{XY}(\tau) = \frac{1}{M}\sum_{i=0}^{M} x(i)y(i-\tau) \tag{5}$$

Kde M je délka posloupnosti, x je první posloupnost, y je druhá posloupnost.

Pokud je potřeba zhodnotit sílu vazby mezi výsledky, lze použit Pearsonův korelační koeficient, který se počítá podle vzorce:

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{6}$$

Jako X a Y jsou označené posloupnosti. Pruhem jsou označené jejích průměrná hodnota.

Teoretický korelace mezi zcela nezávislými signály by se měla rovnat nule. Ale tomu tak není, protože korelační koeficient je odhadem. Nejčastěji hodnoty jsou buď kladné nebo záporné, zároveň absolutní hodnota může nabývat maximálně hodnoty 1. V případě kladné hodnoty mezi signály je nějaká souvislost, v případě záporné hodnoty je také závislost, ale reverzní. Významná tato závislost ale je jenom nad nějakým prahem.

Vzdálenostní metriky

Kromě korelace jsou i další způsoby vyjádřit podobnost (příp. nepodobnost) dvou číslicových signálů. Jednou z takových metod je výpočet vzdálenosti mezi signály, přičemž existují více druhů vzdálenosti. Před popisem těchto druhů samotných, třeba definovat pojem vzdálenostní metriky.

Pokud je definovaná nějaká množina bodů, jinak nazývaná "prostor", vzdálenostní metrika pro tento prostor je funkce *d(x, y)*. Argumenty *x* a *y* jsou dva body v prostoru a výsledkem je vždy reálné číslo, které splňuje následující axiómy:

1. Vzdálenost nemůže být záporná.
2. Vzdálenost se rovná nule, pokud body mají stejnou hodnotu: *x=y*.
3. Vzdálenost je symetrická veličina: *d(x, y)= d(y, x)*.
4. Trojúhelníková nerovnost: $d(x, y) \leq d(x, z) + d(z, y)$. Při splnění této axiómy, vzdálenostní metrika bude popisovat nejkratší vzdálenost bodu x od bodu y. [19]

V případě porovnaní dvou signálů argument *x* označuje první signál a argument *y* označuje druhy signál. Oba signály musejí být stejné délky, proto podobně jako v případě korelačního

`

koeficientu, při vypočítávání vzdálenosti mezi malým úsekem a celým genomem třeba posouvat kratší signál vůči delšímu a vypočítávat vzdálenost překrývajících se časti.

Dole budou definované dvě nejvýznamnější metrické vzdálenosti a jejích modifikace. Tyto dvě metody jsou : Euklidovská a Hammingova vzdálenost.

Nejpoužívanější vzdálenostní metrikou (dále VM) je Euklidovská vzdálenost, definovaná vztahem (2).

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (7)$$

Kde n je délka srovnávaných vektorů *x* a *y*.

Používá se také Euklidovská vzdálenost normalizována, vážená, s *r*-tou odmocninou, kde *r* je jakékoliv konstanta a také Euklidovská vzdálenost na druhou.

Další VM je Hammingova metrika: viz vzorec (3).

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sum_{i=1}^{n} |x_i - y_i| \qquad (8)$$

Hammingova vzdálenost je nejmenší počet změn po kterým vektor *x* byl přeměněn do vektoru *y*. Pro různé typy mapování může být výsledek interpretován různě. Tento typ VM může být také normalizován. [20]

Minkovského metrika je třetím typem. Počítá se podle vzorce (4)

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \left[\sum_{i=1}^{n} |x_i - y_i|^m\right]^{\frac{1}{m}} \qquad (9)$$

Kde m je jakákoliv konstanta.

Minkovského vzdálenost je v podstatě sloučením Euklidovské a Hammingove metriky. Třeba poznamenat, že velikost mocniny musí být volená podle míry podobnosti dvou signálů.

Canberrská metrika je definovaná vztahem (5). Je to citlivější varianta Hammingove vzdálenosti.

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \qquad (10)$$

`

Koeficient divergence je obdobná metrika jako Canberrská. Zakládá se na euklidovské vzdálenosti a je normalizována. Viz vztah (6)

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x_i - y_i|}{|x_i| + |y_i|}\right)^2} \qquad (11)$$

Je patrné, že podle vztahů (5) a (6) nelze vypočítat vzdálenost dvou nulových bodů (objeví se nula v jmenovateli). Proto tyto případy je potřeba řešit před počítáním, např. doplnit hodnotu 0 na těchto pozicích. [21]

Výše popsané metody vypočítávají vzdálenost signálů, to znamená stupeň nepodobnosti. Čím větší je hodnota, tím méně podobné jsou sekvence. Proto tentokrát po výpočtu vzdálenosti bude nás zajímat nejmenší hodnota. Popsané metody jsou výpočetně jednoduché, avšak není jisté, zda můžou odhalit příbuzenství genu. Pokud tyto metody budou použité pro vyhledávání homologních genů, musí být určen práh, nad kterým podobenství je spíš náhodné.

`

4 VLASTNÍ METODY VYHLEDÁVÁNÍ GENŮ

V následující kapitole budou popsané ukázky vyhledávání podobnosti (příp. nepodobnosti) genů pomocí korelace, euklidovské vzdálenosti a canberrské vzdálenosti signálů. Bude se srovnávat cely genom a jeho úsek (CDS úsek). Ve všech případech byl použit stejný genom a stejný jeho CDS úsek, aby výsledky byli porovnatelné mezi sebou.

Gen byl převzat z databáze GenBank na webové stránce NCBI (angl. The National Center for Biotechnology Information). GenBank databáze je součástí mezinárodní organizace (angl. International Nucleotide Sequence Database Collaboration) spojující databáze Japonska, Europy a USA.

V dnešní době GenBank obsahuje nukleotidové sekvence více než 260 000 popsaných druhů organismů. Může je nahrávat vědci po celém světě, pokud mají přístup k speciálním programům: Sequin a BankIt. Je to jedna z nejvýznamnějších databází, avšak i v ní lze potkat chyby.

Při testování všech navřených programů byl použit kompletní mitochondriální genom myší domácí (*Mus musculus*) s identifikátorem AP013031.1, podle kterého lze tuto sekvenci najít na stránkách NCBI. Kruhová sekvence DNA má délku 16300 bp a obsahuje 13 CDS úseku. Pro ukázky kódu byl použit CDS úsek kódující cytochrom b. Je v genomu na pozicích 14146 až 15289.

4.1 Vyhledávání s využitím korelace

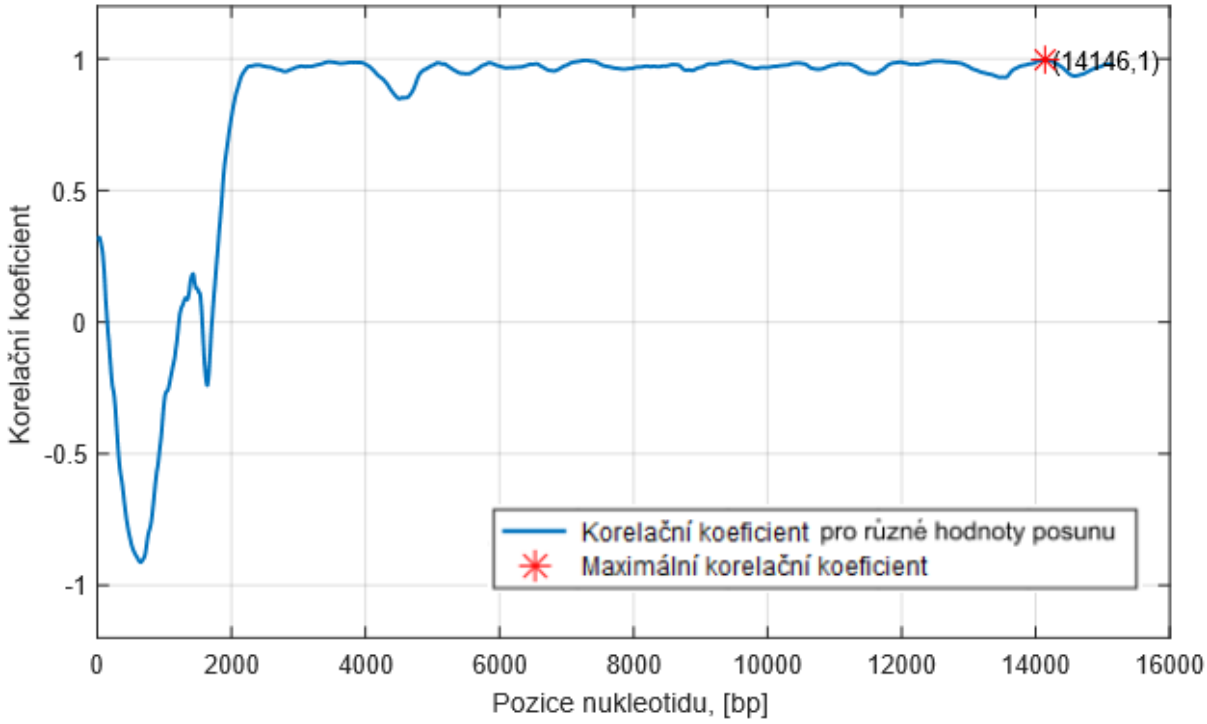
Základem je výpočet korelačního koeficientu podle vzorce (6) mezi CDS úsekem a celým genomem. Pro obě numerické reprezentace je postup stejný:

1. Sekvence se převádí na číslicový signál.
2. Signál reprezentující genom a jeho úsek na pozicích 14146 až 15289 se zarovnávají od začátku genomu.
3. Vypočítá se normalizovaný korelační koeficient pro překrývající se vektory.
4. Genom se posune o jeden nukleotid tak, že teď se překrývají CDS úsek a část genomu od 2. nukleotidu do (N+1) nukleotidu, kde N je délka CDS úseku.
5. Posun a výpočet koeficientu se opakuje do konce genomu.
6. Pozice maximálního koeficientu je začátkem hledaného CDS úseku v genomu.

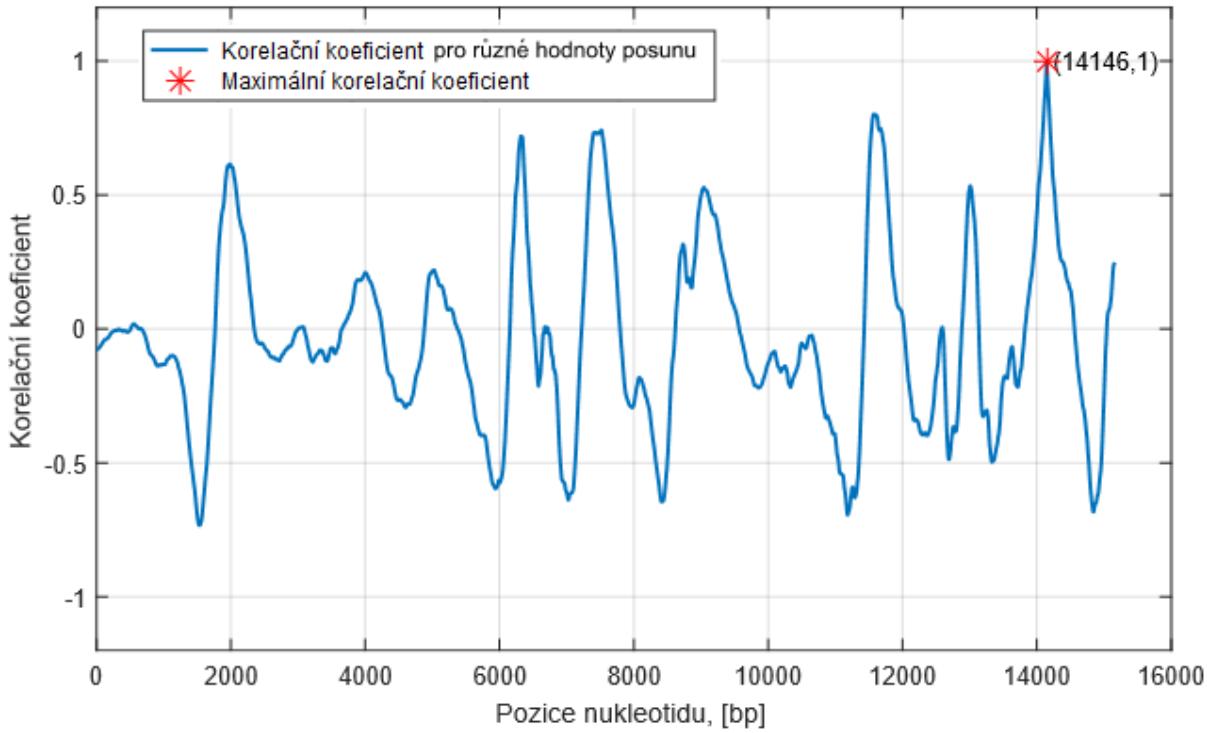
Pokud na x-ovou osu bude umístěn posun (neboli začátek zarovnávání) a na y-ovou osu odpovídající korelační koeficient, dostaneme následující křivky: pro kumulovanou fázi (viz Obr. 4.1) a rozbalenou fázi (viz Obr. 4.2).

Je hned patrné, že kumulovaná fáze má lokální maxima velmi blízké k hodnotě 1. Rozbalená fáze má pouze několik píků kolem hodnoty 0,8. V případě vyhledávání úplně identického úseku v genomu, žádný problém spojený s těmito lokálními píky nejsou. Obě reprezentace korelují s CDS úsekem přesně na pozici jeho umístěni v genomu. Avšak v případě méně podobných sekvencí mohla by nastat chyba (falešní detekce) a správný homologní úsek by byl zanedbán.

`



Obr. 4.1 – Korelační koeficienty pro různé pozice hledaného genu vůči celému genomu (kumulovaná fáze)



Obr. 4.2 - Korelační koeficienty pro různé pozice hledaného genu vůči celému genomu (rozbalená fáze)

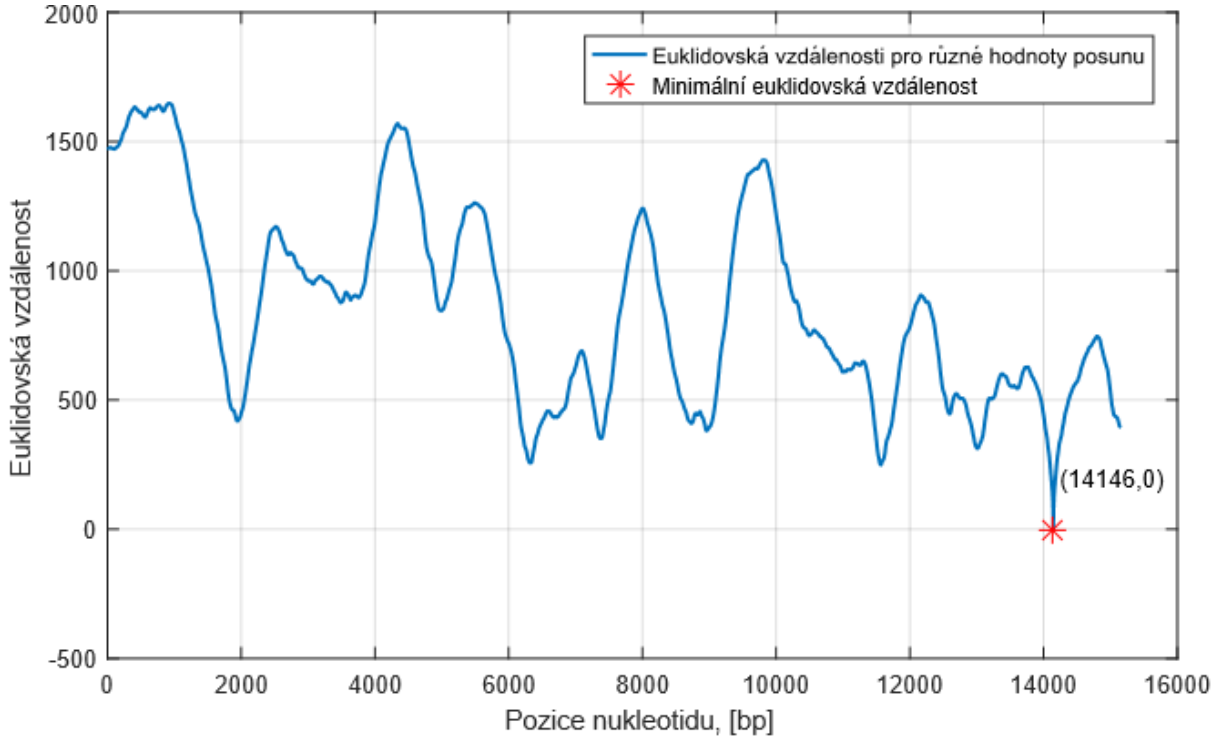
4.2 Vyhledávání s využitím euklidovské vzdálenosti

Postup pro vyhledávání s využitím euklidovské vzdálenosti je velmi podobný:

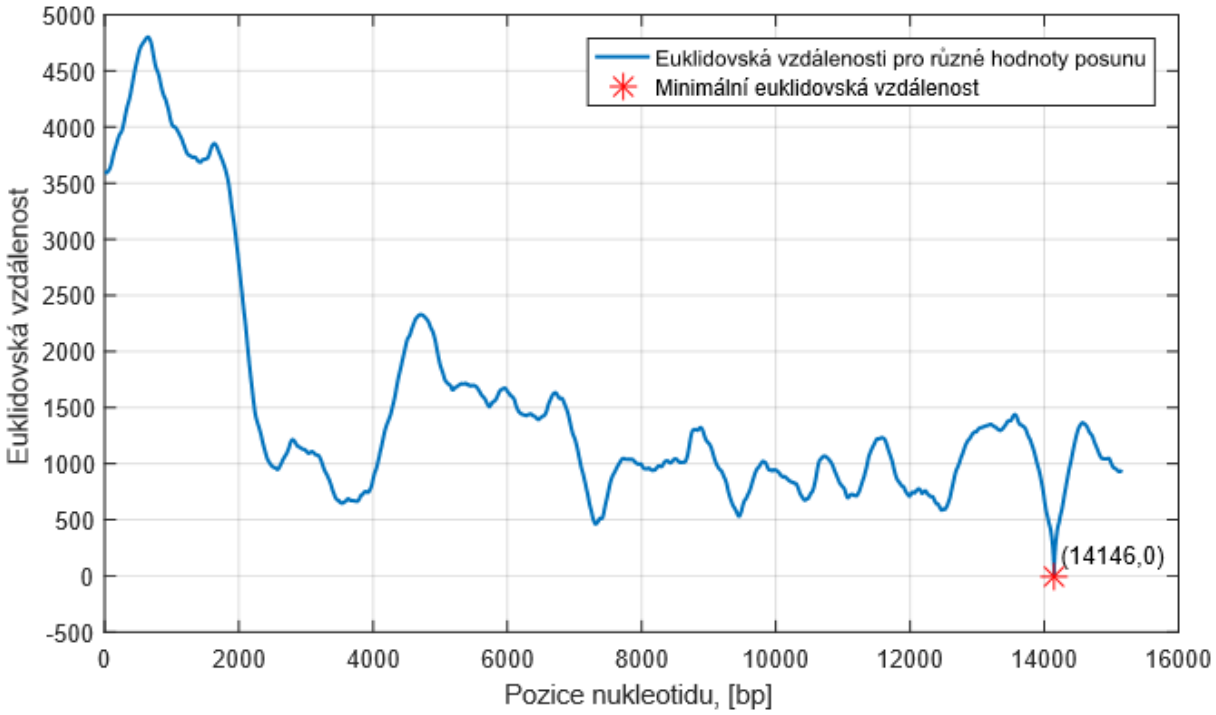
1. Sekvence se převádí na číslicový signál.
2. Signál reprezentující genom a jeho úsek na pozicích 14146 až 15289 se zarovnávají od začátku genomu.
3. Vypočítá se euklidovská vzdálenost pro překrývající se vektory.
4. Genom se posune o jeden nukleotid tak, že teď se překrývají CDS úsek a část genomu od 2. nukleotidu do (N+1) nukleotidu, kde N je délka CDS úseku.
5. Posun a výpočet vzdálenosti se opakuje do konce genomu.
6. Pozice minimální vzdálenosti je začátkem hledaného CDS úseku v genomu.

Tentokrát třeba hledat minimální hodnotu vzdálenost, protože na rozdíl od korelačního koeficientu, který je mírou podobností, euklidovská vzdálenost je mírou nepodobností. Tato metoda také zvládla najít správnou pozici začátku CDS úseku.

Obě metody mají velký rozsah hodnot vzdálenosti. Pro rozbalenou fázi (obr. 4.3) je rozsah kolem 0 až 1650. Pro kumulovanou(obr. obr. 4.4) je to kolem 0 až 4830. Třeba také poznamenat, že ani kumulovaná ani rozbalená fáze neměla ve výsledku 0. Bylo to číslo velmi malé, které se blíží nule.



Obr. 4.3 - Euklidovská vzdálenost pro různé pozice hledaného genu vůči celému genomu (rozbalená fáze)



Obr. 4.4 - Euklidovská vzdálenost pro různé pozice hledaného genu vůči celému genomu (kumulovaná fáze)

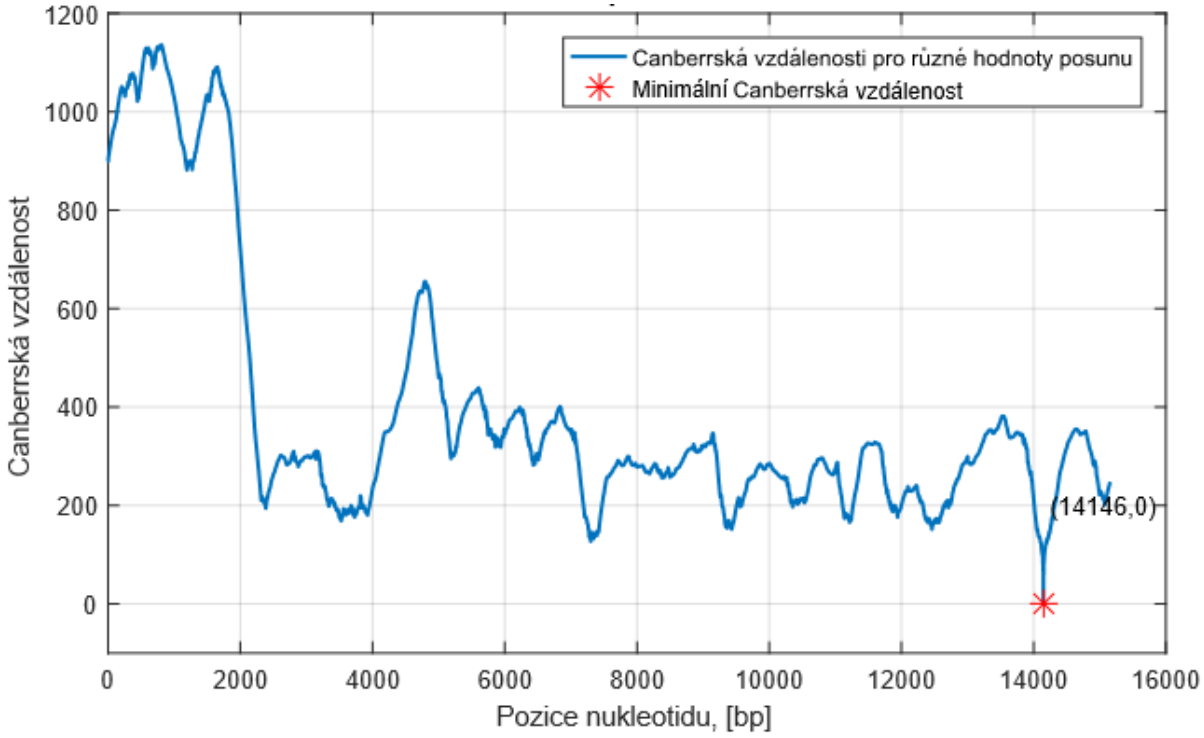
4.3 Vyhledávání s využitím canberrské vzdálenosti

Postup vyhledávání je stejný jako v případě euklidovské vzdálenosti, avšak se vypočítává canberrská vzdálenost.

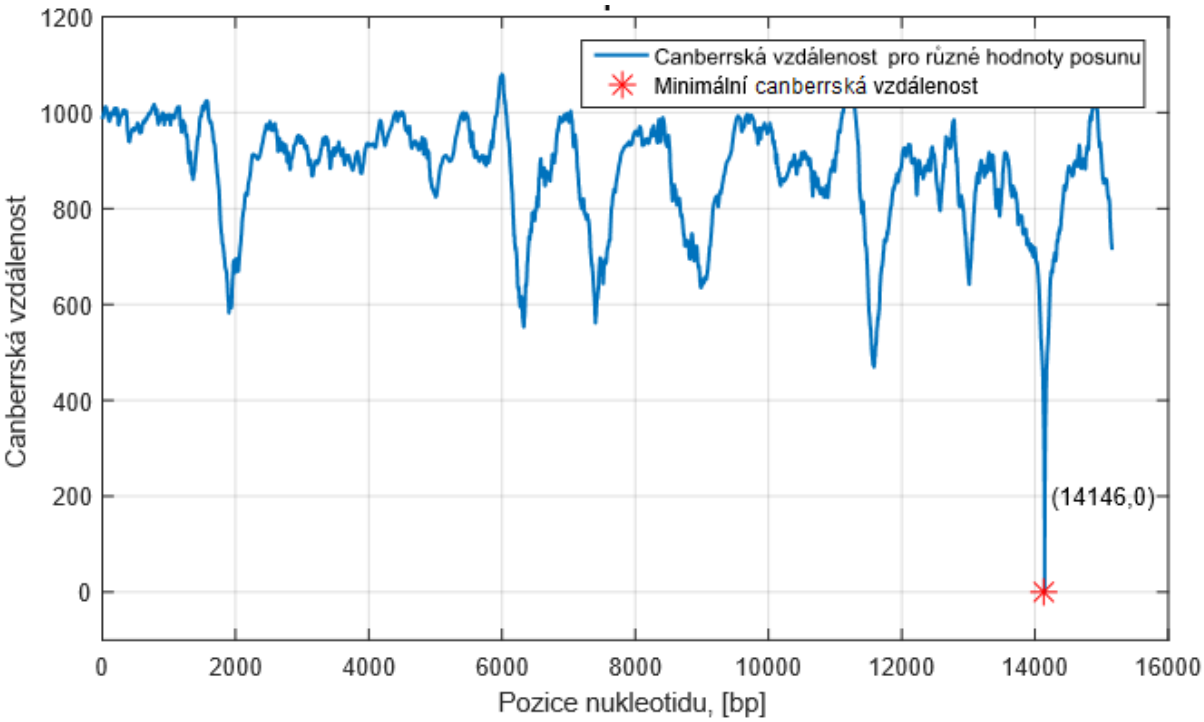
Z výsledků je patrné, že rozsah je menší, než u euklidovské metriky. Pro rozbalenou fázi (obr. 4.6) je rozsah kolem 0 až 1120. Pro kumulovanou(obr. obr. 4.5) je to kolem 0 až 1150.

Rozbalená fáze má nejmenší lokální minima ze všech použitých metod, což je pro vyhledávání nejvhodnější. Signál připomíná zašuměné kolísání kolem hodnoty 1000 s záporným skokem v místě, kde je v genomu umístěn CDS úsek.

Kumulovaná fáze naopak ve všech případech nevypadá jako nejvhodnější metoda. Z obrázků je vidět, že na začátku jsou hodnoty nejvíce rozlišné, když žádná z metod při použití rozbalené fáze tuto skutečnost nepotvrdila. Možným důvodem je nevhodně vybraný zlomek ve vzorci (2).



Obr. 4.5 - Canberrská vzdálenost pro různé posuny (kumulovaná fáze)



Obr. 4.6 - Canberrská vzdálenost pro různé posuny (rozbalená fáze)

5 VLASTNÍ METODY VYHLEDÁVÁNÍ HOMOLOGNÍCH GENŮ

Minulá kapitola se zabývala metody vyhledaní určitého úseku genu v celém genomu. V následující kapitole jsou uvedené výsledky vyhledání genů homologních, a to v genomech více či méně příbuzných organizmů.

Treponema pallidum a *Treponema paraluiscuniculi* jsou bakterie, DNA kterých je tvořena jedním kruhovým chromozomem. Pro testování programu byli použité následující poddruhy organismu *Treponema pallidum*: *Treponema pallidum subsp. pertenue str. Samoa D* (délka 1139330 nukleotidů), *Treponema pallidum subsp. pallidum str. Nichols* (délka 1139633 nukleotidů), *Treponema pallidum subsp. pallidum str. Mexico A* (délka 1140038 nukleotidů), *Treponema pallidum subsp. pertenue str. CDC2* (délka 1139744 nukleotidů) a také poddrůh *Treponema paraluiscuniculi Cuniculi A* (délka 1133390 nukleotidů). Data jsou převzata z databáze NCBI.

V každém z výše uvedených genomu budou prohledané pět genu organismu *Treponema pallidum subsp. pertenue str. Samoa D.* Jsou to geny kódující: tetrakopeptid A, tetrakopeptid C, tetrakopeptid D, tetrakopeptid E a (dle předpokladu) protein vnější membrány. Kromě genu tetrakopeptidu C, jsou geny umístěné na komplementárním vlákně.

Vybrané geny organismu *Samoa D* jsou více či méně podobné analogickým genům u homologních organismů. Pro některé geny lze snadno interpretovat výsledky, u jiných lze jen předpokládat, zda program našel správný úsek. Na základě prvních se nastavuje práh vzdáleností, podle kterého se bude rozhodovat, jestli je gen homologní, v případě druhých se tento práh jen používá.

Na základě výsledků v předcházející kapitoly bylo rozhodnuto použit 4 numerické reprezentace: rozbalenou fázi a denzitní vektory s různými délkami okna. Také bylo rozhodnuto otestovat jen euklidovskou a canberrskou vzdáleností jako způsob vyhledání, protože se ukázalo, že korelace není dostatečně vhodná metoda.

5.1 Popis programu

Vlastní řešení je vypracované na základě programů popsaných v minulé kapitole. Funkce vypočítávající canberrskou, euklidovskou vzdálenosti a denzitní vektory byli převzatý beze změn, avšak jako funkce rozbalené fáze byla použitá jednodušší varianta výpočtu (viz tab. 3.2). Důvodem je časová náročnost výpočtu podle vzorce (3). Na rozsáhlejších datech tento typ výpočtu rozbalené fáze trval několikanásobně víc, přičemž použitý výpočet byl dokonce výhodnější z důvodu snadného odvození signálu komplementárního vlákna.

Postup:

1. Program načitá vstupní data (genom a 5 genu).
2. Vše signály jsou převedené do číslicového tvaru (rozbalená fáze a denzitní vektory C+A s oknem 3,5,17).
3. Jsou získané signály pro komplementární vlákno všech pěti genů. V případě rozbalené fáze byl použit převod pomocí vztahu (5), v případě denzitních vektorů byla sekvence

`

převedená na komplementární vlákno a následně přepočítaná na číslicový signál stejným způsobem jako kódující vlákno.

4. Počítá se euklidovská a canberrská vzdálenost pro kódující a komplementární vlákna genu.
5. Minimální hodnota každé vzdáleností je uložená, stejně jako pozice, na které hodnota je minimální a celý signál vypočítané vzdálenosti.

Na základě toho, jaký signál má nejmenší minimální hodnotu, rozhoduje se, zda je homologní gen přítomen na kódujícím nebo na komplementárním vlákně. Tuto operaci by mohl provádět rovněž program, nicméně program ukládá oba typy signálu a není potřeba vyznačovat na jakém signálu je gen s největší pravděpodobností přítomen: tato informace níž bude znázorněna v grafech.

Zdrojem pěti vybraných genů je poddruh *Treponema pallidum Samoa D*. V tomto genomu, stejně jak i v ostatních se budou vyhledávat hodnoty vzdálenosti. Výsledky mohou být použitý jako vzor velmi podobných homologních genu a dají nám představu o rozsahu hodnot, minimální možné hodnotě vzdáleností atd.

5.2 Výsledky vyhledaní homologních genů

V této kapitole budou shrnuté a diskutované výsledky pro každý genom zvlášť.

Pro názornost budou níž uvedené tabulky výsledků a grafy. Proto je potřeba uvést co znamenají některé zkratky použité v tabulkách:

RF: rozbalená fáze.
DV(5): denzitní vektory s oknem 5.
Gen 1.: tetrakopeptid A
Gen 2.: tetrakopeptid C
Gen 3.: tetrakopeptid D
Gen 4.: tetrakopeptid E
Gen 5.: protein vnější membrány (dle předpokladu)
Gen 4.(komp.): komplementární vlákno tetrakopeptidu E
Šedě jsou označené buňky s minimální hodnotou vzdáleností: buď kódující vlákno a/nebo komplementární.

Hodnoty vzdáleností v obrázcích jsou zaokrouhlené. Vše grafy lze najit v příloze.

Genom *treponema pallidum subsp. pertenue str. Samoa D*

Podle detekovaných pozic (viz tab. 5.1) je vidět, že vše metody nalezli správné umístění genu. Hodnoty vzdálenosti u různých metod se liší.

U rozbalené fáze vzdálenost na správném vlákně je nulová, na opačném dosahuje až 97 pro euklidovskou vzdálenost a až 1032 pro canberrskou vzdálenost.

U denzitních vektorů s použitím okna délkou 3 jsou vzdálenosti na správném vlákně menší než 1 pro euklidovskou vzdálenost a menší než 4 pro canberrskou. Pro opačné vlákno dosahuje hodnot 16 a 996 pro různé typy vzdálenosti.

`

Denzitní vektory s použitím delších oken hodnoty vzdálenosti zvětšuje jak na správném vlákně, tak i na opačném. Pro správné vlákno euklidovské vzdálenosti je hodnota pro okno 5 stejně menší než 1. Avšak pro okno délkou 17 se rovná 1,5 euklidovské vzdálenosti a 134 canberrské.

Grafy 2., 3. a 4. genu mají i lokální minima, na rozdíl od grafů 1. a 5. genu. Mohlo by to znamenat, že v genomu je několik homologů pro tyto geny. V toho plyne, že práh pro rozlišení homologních a nehomologních genu třeba nastavovat i podle toho, v jaké míře musejí hledané geny být podobné a kolik maximálně výsledků požadujeme. (Méně podobných bude nalezeno víc). Daný program je nastaven tak, aby hledal jeden nejpodobnější homologní gen (v případě dvou stejných hodnot vzdálenosti se používá ta, která je na pozici s menším číslem).

Je také patrné, že v případě grafu 1. -3. a 5. genu hodnota na komplementárním vlákně je mnohem menší a lze snadno rozlišit, zda se gen nachází na kódujícím či komplementárním vlákně.

V Obr. 5.6 až 5.20 lze sledovat jak s velikostí použitého okna denzitních vektorů narůstá počet kmitaní signálů canberrské vzdálenosti. Tyto kmity nemají vliv na výsledek, avšak u hodně odlišných sekvenci by mohli způsobit chybu.

Práh pro euklidovskou vzdálenost by mohl být nastaven na hodnotu 40, 8, 8 a 4(rozbalená fáze, DV s oknem 3,5,17). Pro canberrskou vzdálenost na hodnotu 400 (pro vše reprezentaci).

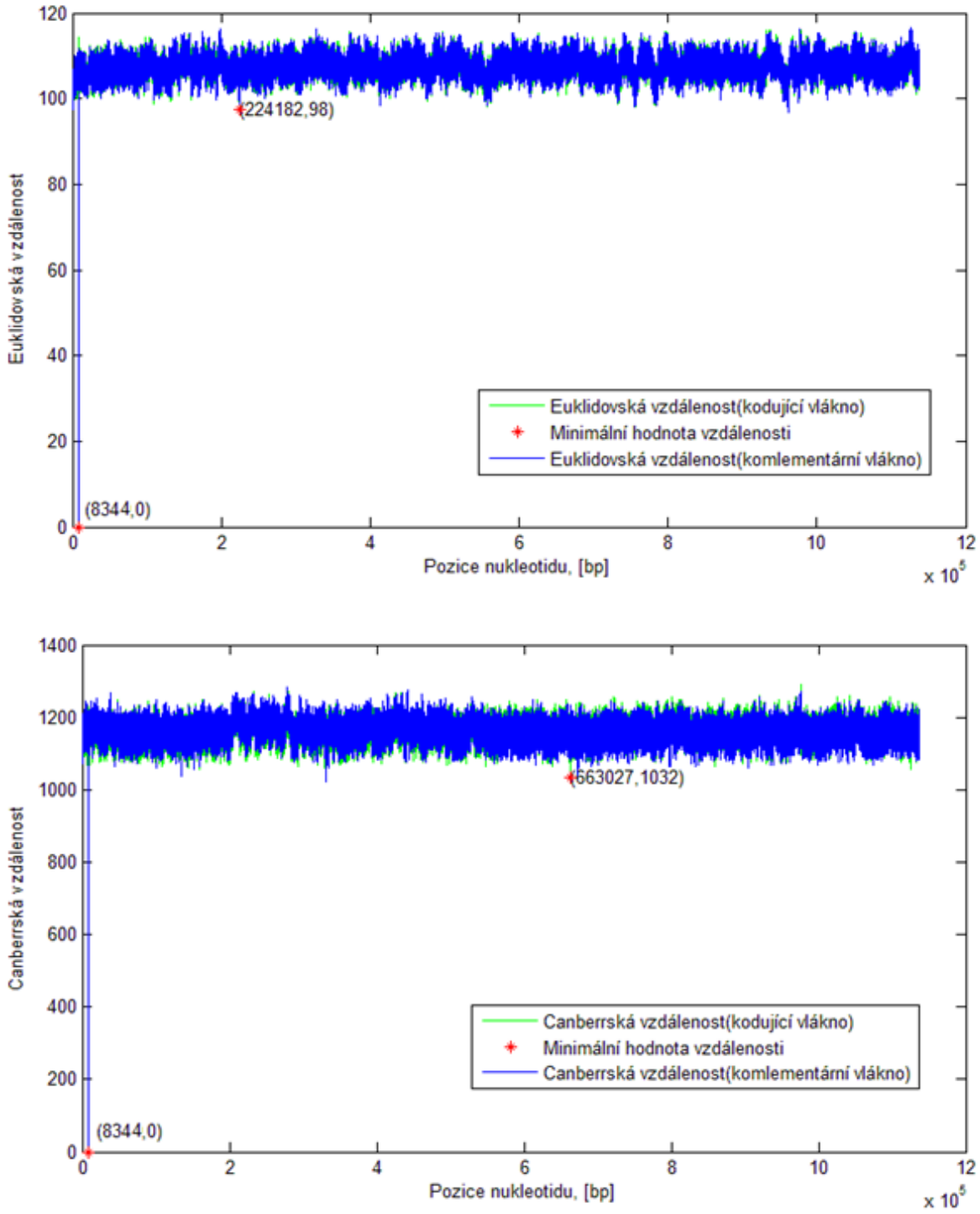
Tabulka 5.1 – Výsledky pro treponema pallidum subsp. pertenue str. Samoa D

Minimální canberrská vzdálenost				**Skutečná pozice**	**Gen**
RF	DV(3)	DV (5)	DV (17)		
663027	838949	15166	229697	-	**1.**
8344	8344	8344	8344	8344	**1.(komp.)**
1125424	1125424	1125424	1125421	-	**2.**
134598	134598	134598	134598	134598	**2.(komp.)**
1125430	663132	805108	1125428	-	**3.**
152044	152044	152044	152044	152044	**3.(komp.)**
328720	328720	328720	328720	328720	**4.**
333744	333744	333744	674522	-	**4.(komp.)**
251759	172472	171592	85458	-	**5.**
1054072	1054072	1054072	1054072	1054072	**5.(komp.)**

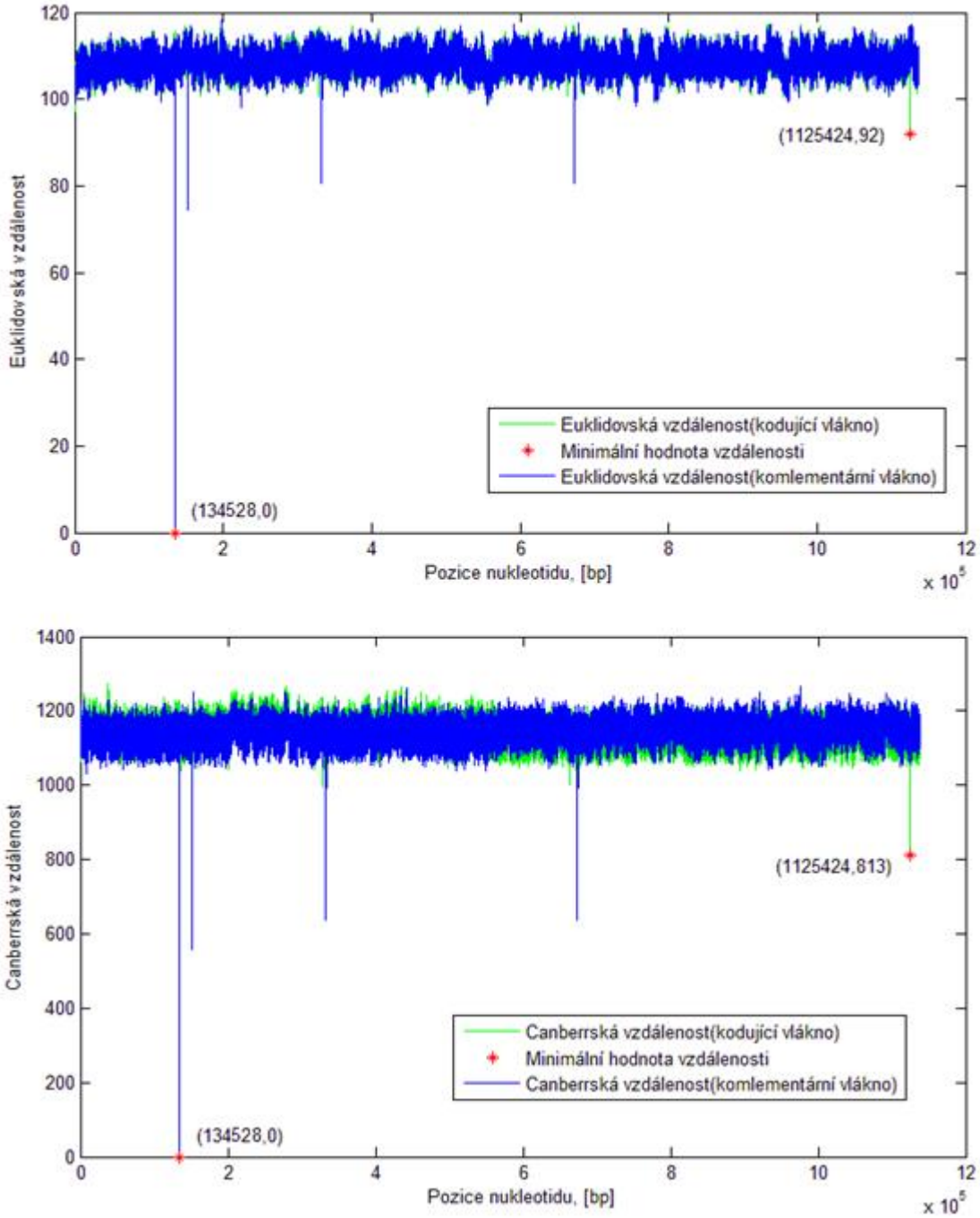
Minimální euklidovská vzdálenost				**Skutečná pozice**	**Gen**
RF	RF	RF	RF		
224182	224182	224182	224182	-	**1.**
8344	8344	8344	8344	8344	**1.(komp.)**
1125424	1125424	1125424	1125424	-	**2.**
134598	134598	134598	134598	134598	**2.(komp.)**
663132	663132	663132	663132	-	**3.**
152044	152044	152044	152044	152044	**3.(komp.)**
328720	328720	328720	328720	328720	**4.**
333744	333744	333744	333744	-	**4.(komp.)**
544	544	544	544	-	**5.**

`

1054072	1054072	1054072	1054072	1054072	**5.(komp.)**

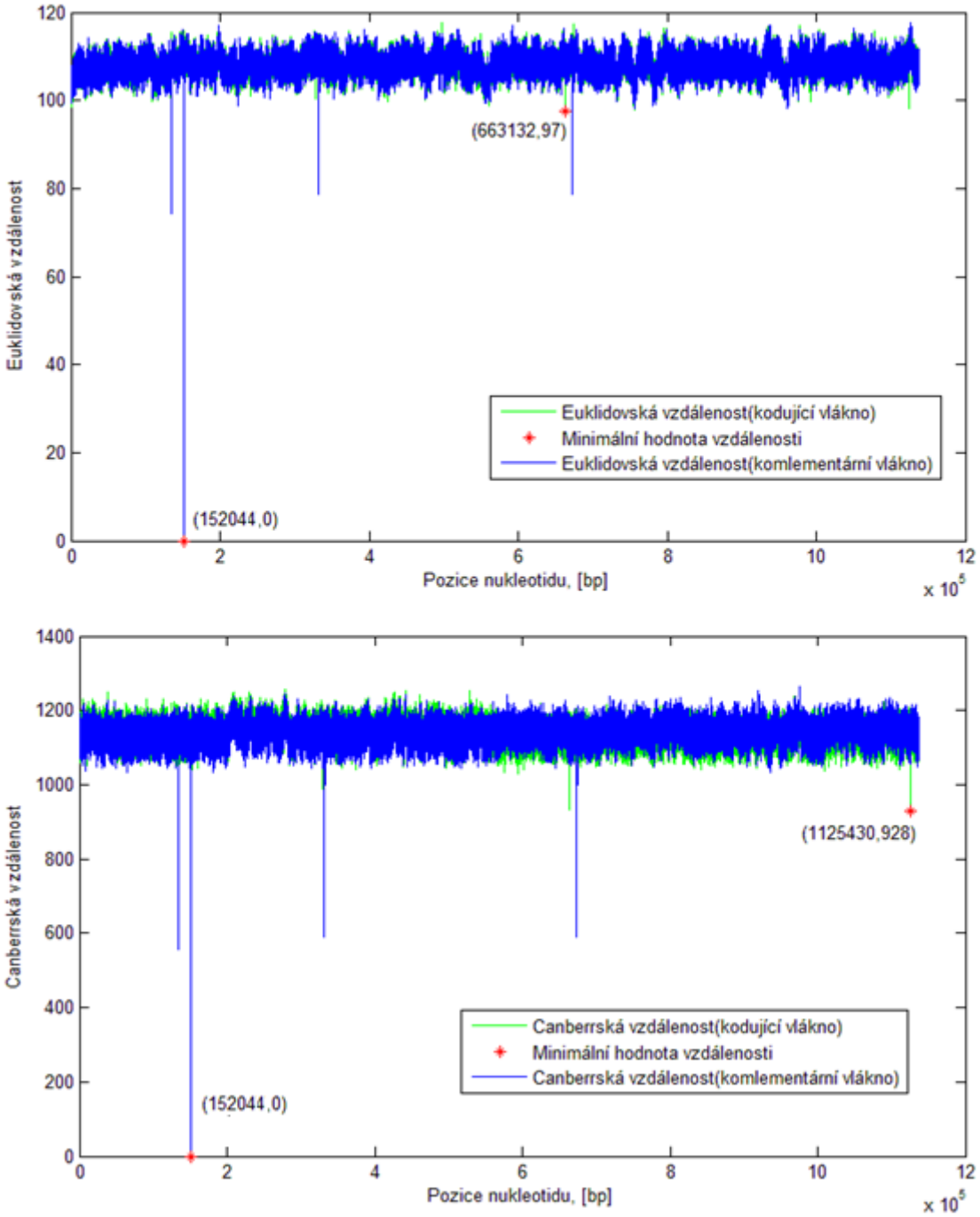


Obr. 5.1 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: rozbalená fáze.

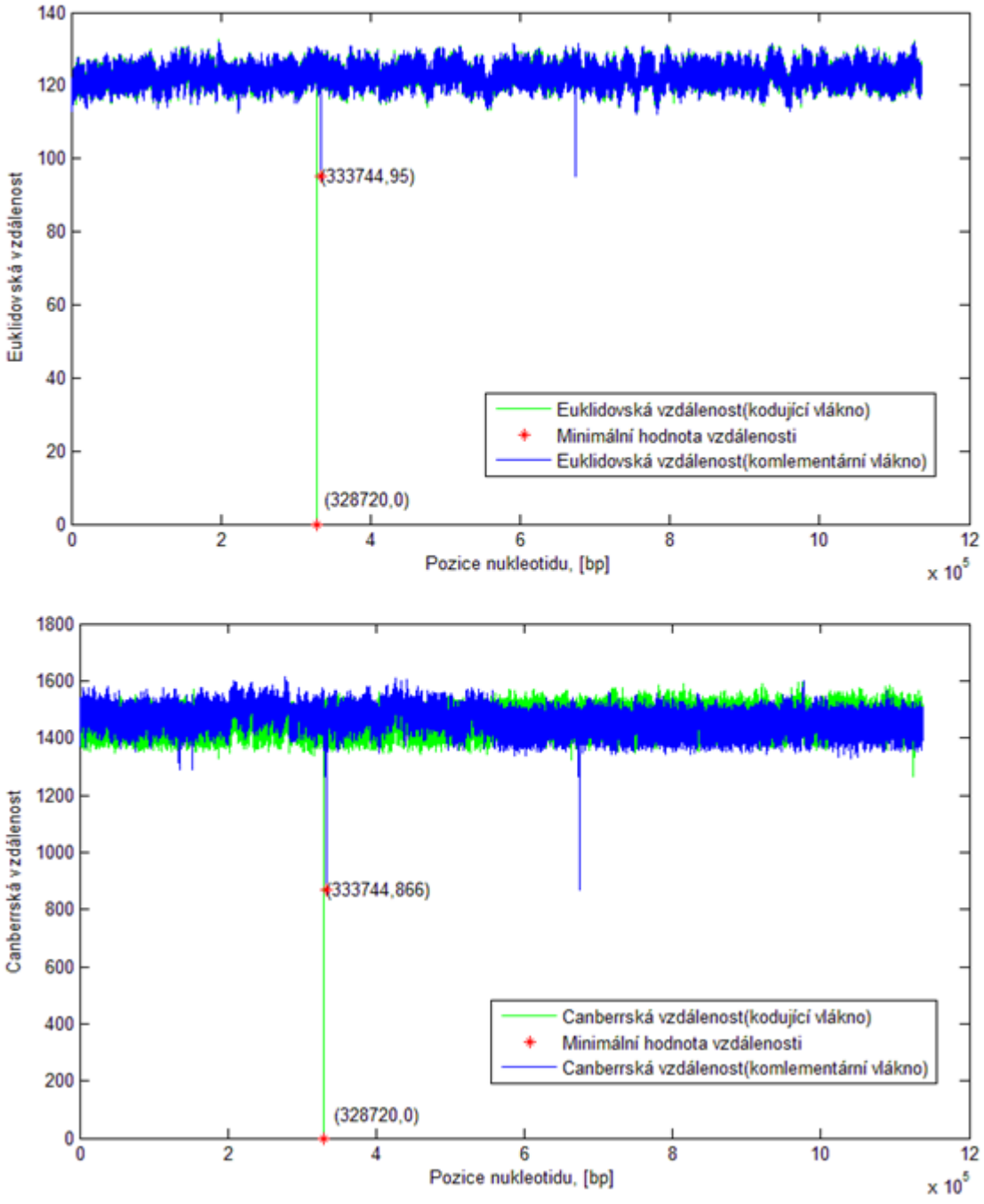


Obr. 5.2 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 2. gen. Numerická reprezentace: rozbalená fáze.

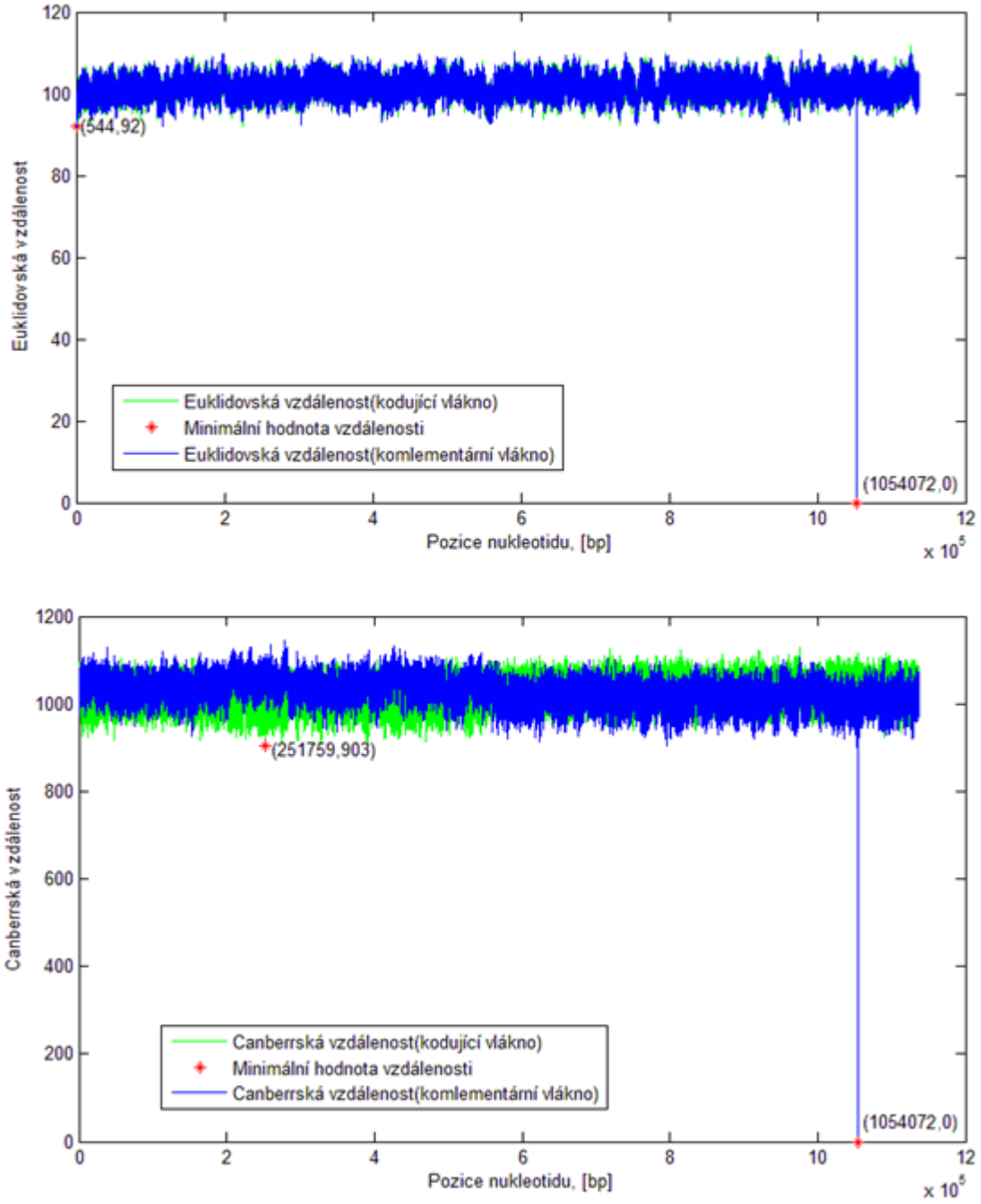
`



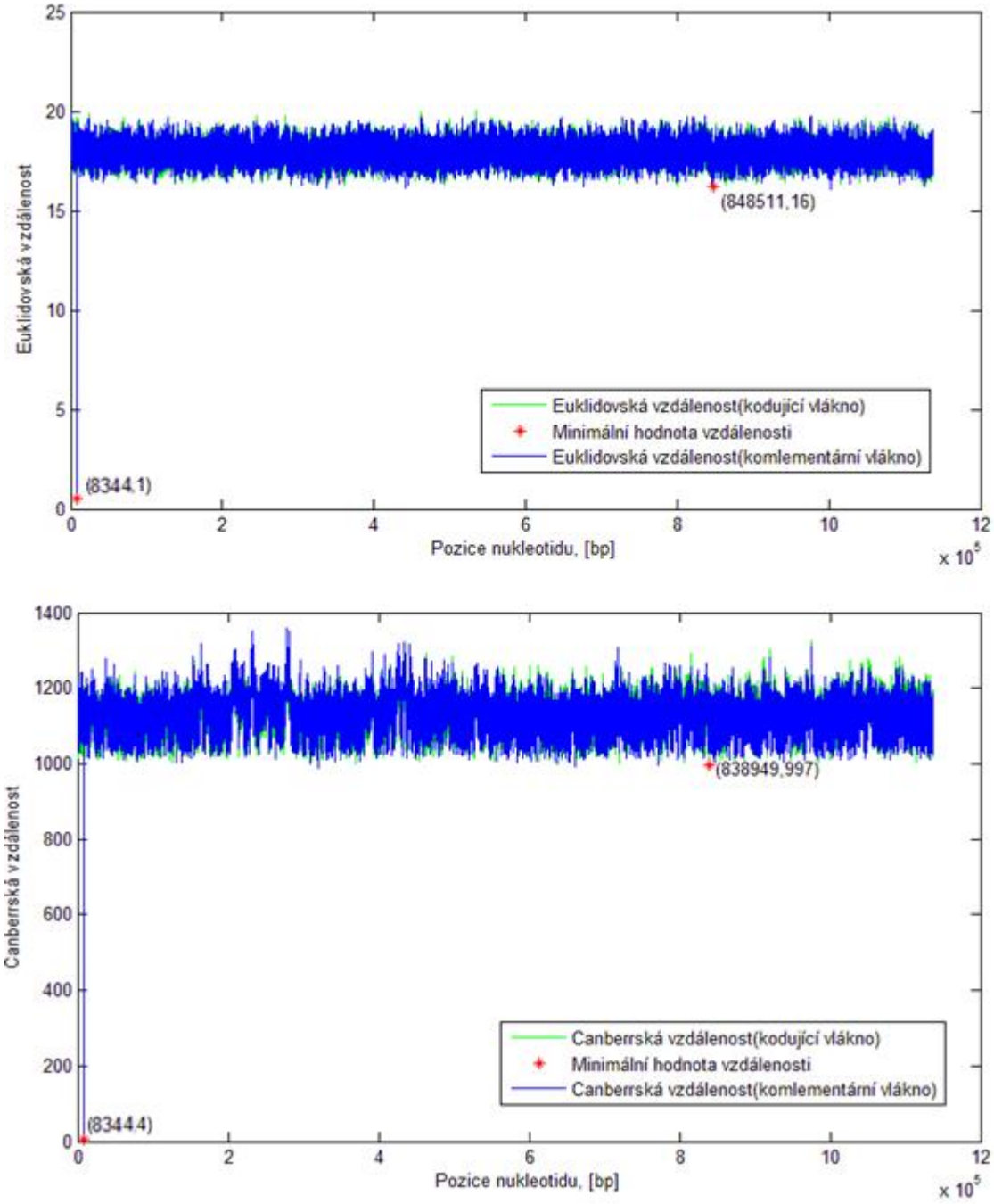
Obr. 5.3 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 3. gen. Numerická reprezentace: rozbalená fáze.



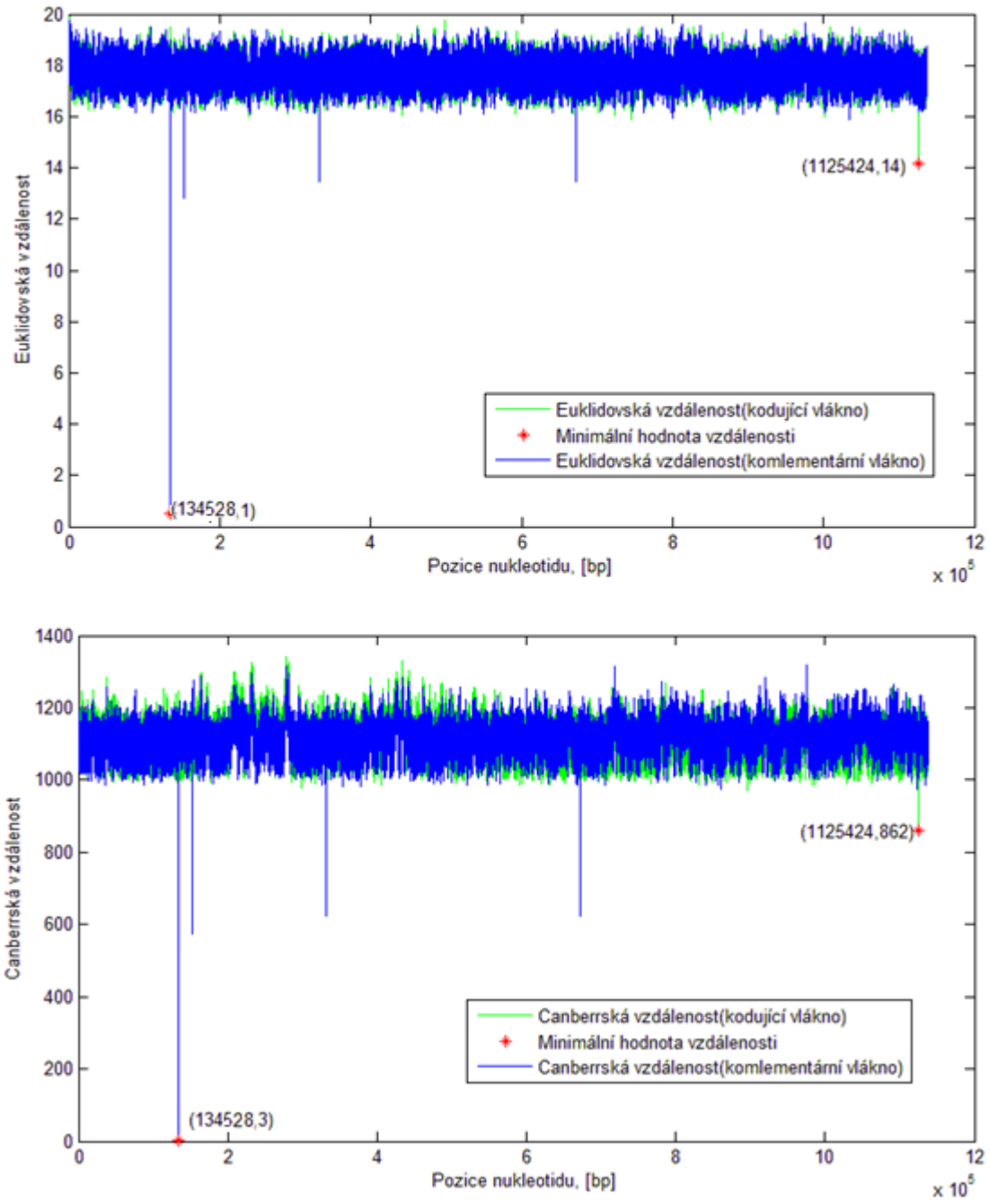
Obr. 5.4 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 4. gen. Numerická reprezentace: rozbalená fáze.



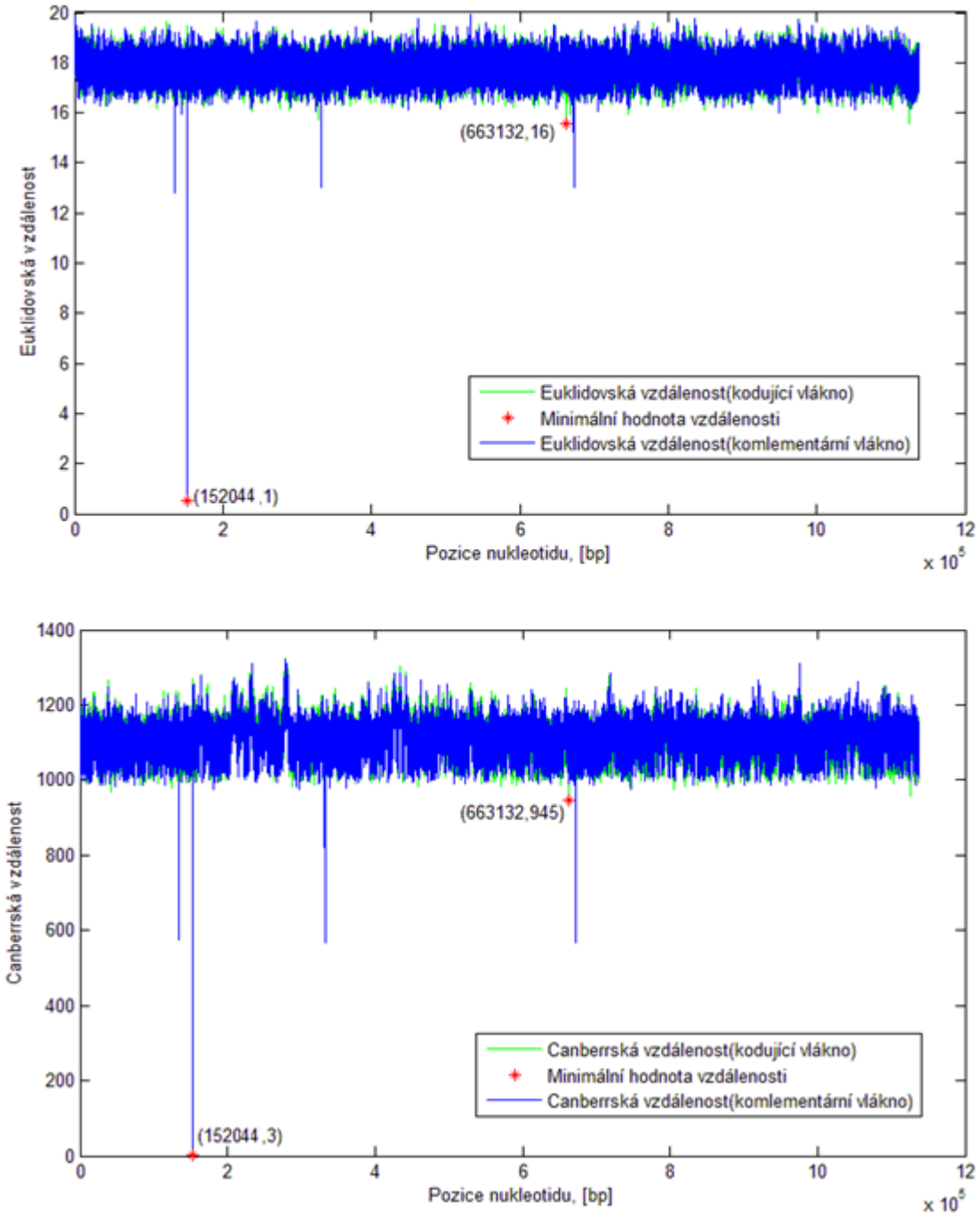
Obr. 5.5 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 5. gen. Numerická reprezentace: rozbalená fáze.



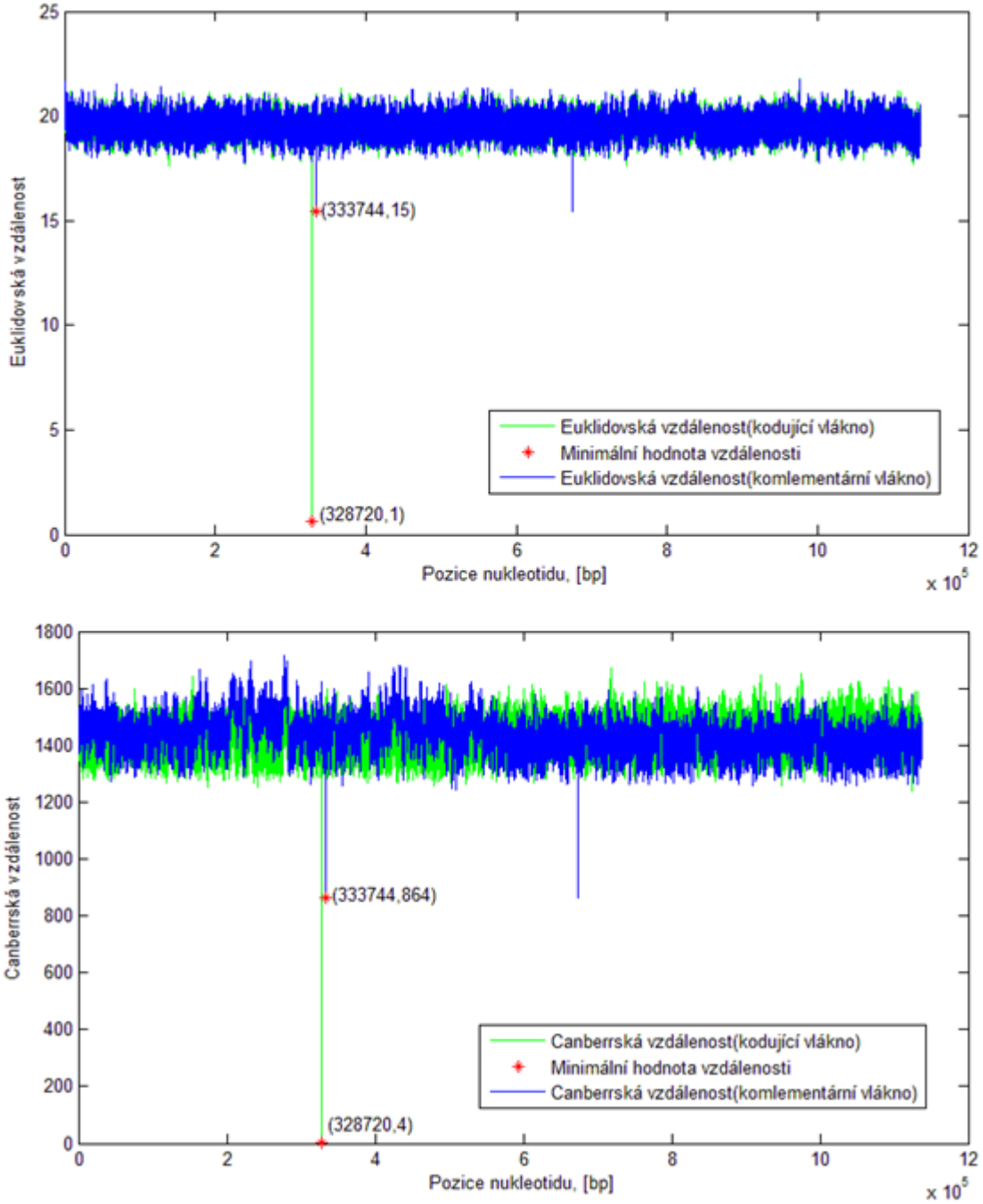
Obr. 5.6 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 3.



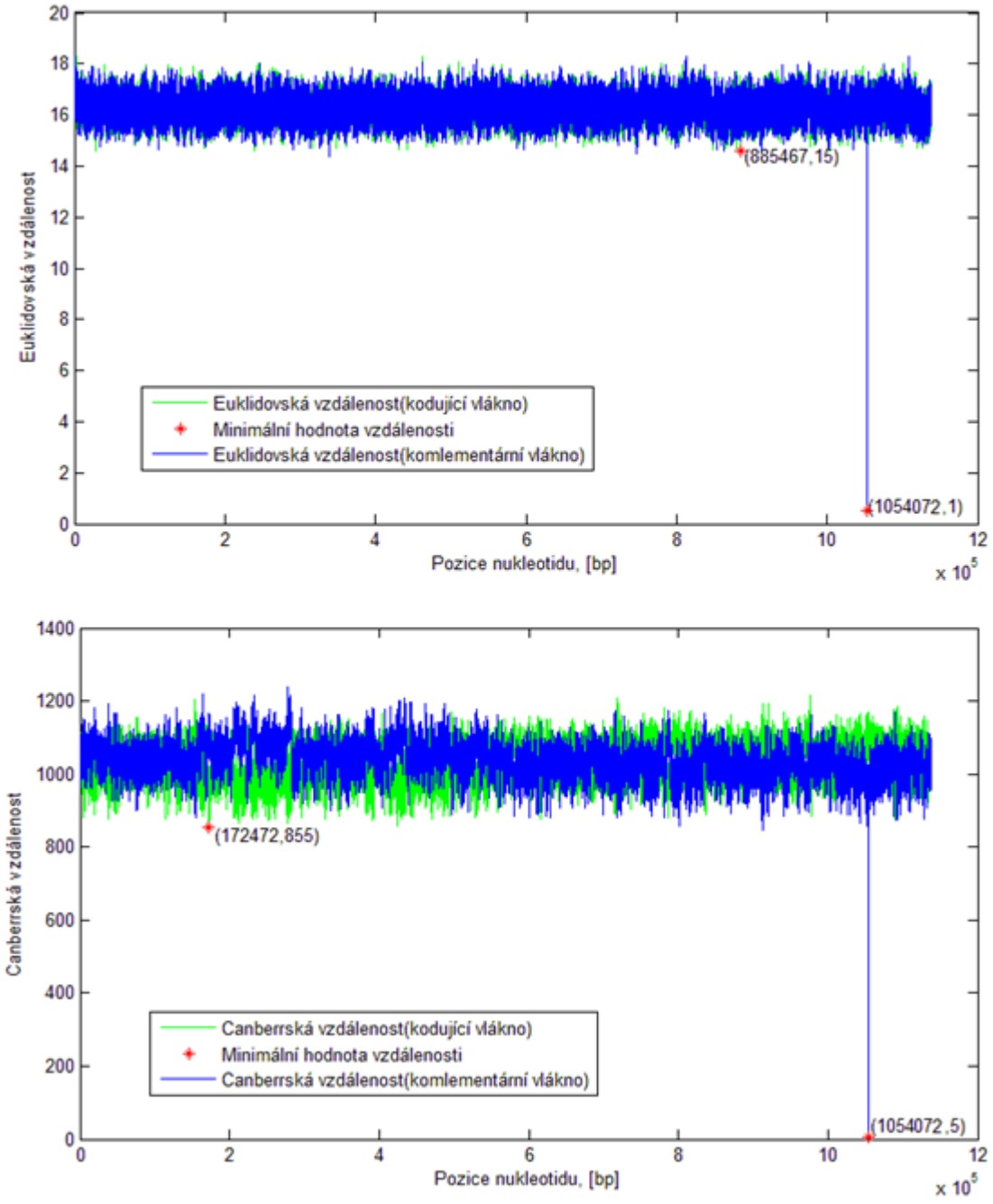
Obr. 5.7 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 2. gen. Numerická reprezentace: denzitní vektory s oknem 3.



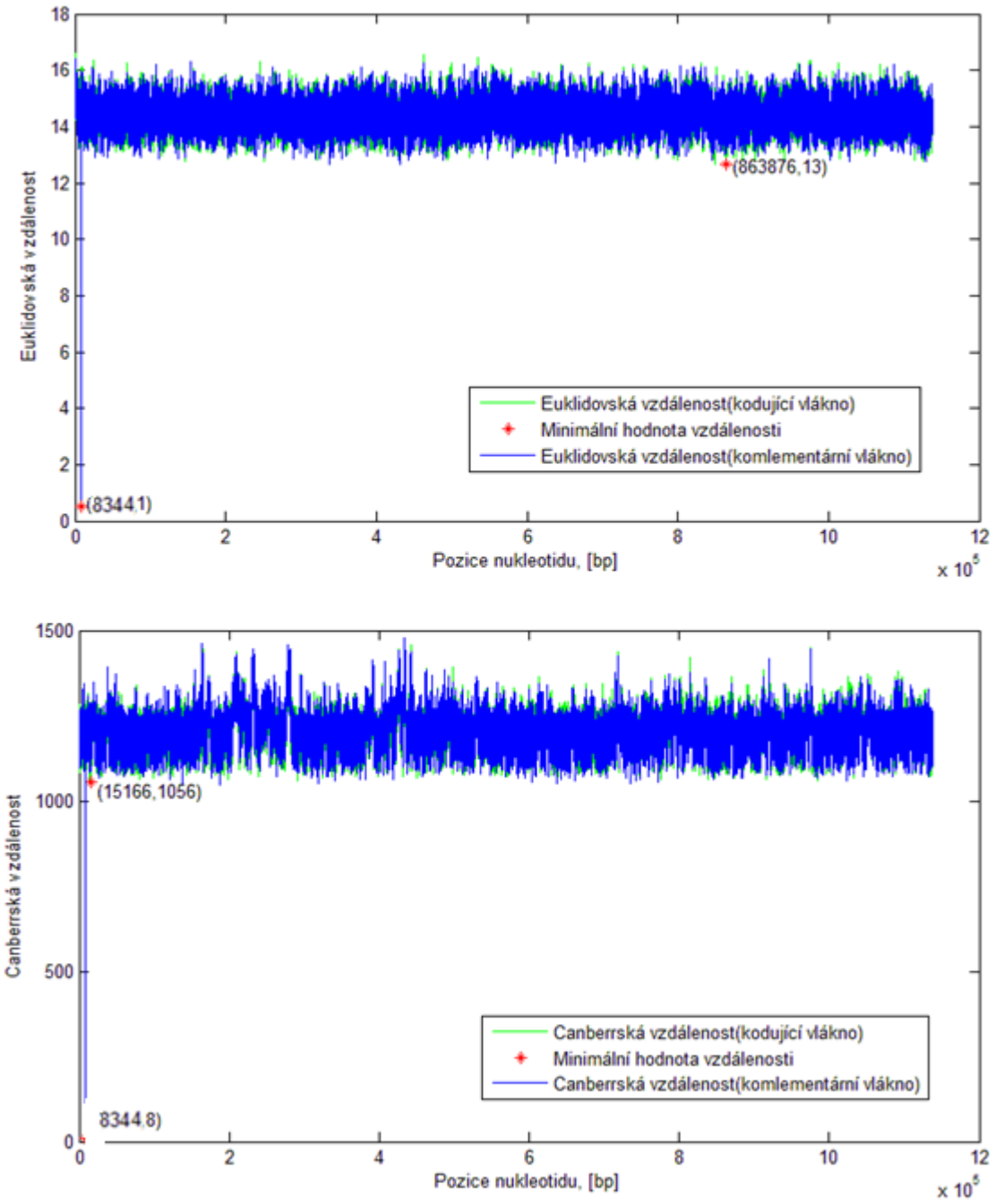
Obr. 5.8 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 3. gen. Numerická reprezentace: denzitní vektory s oknem 3.



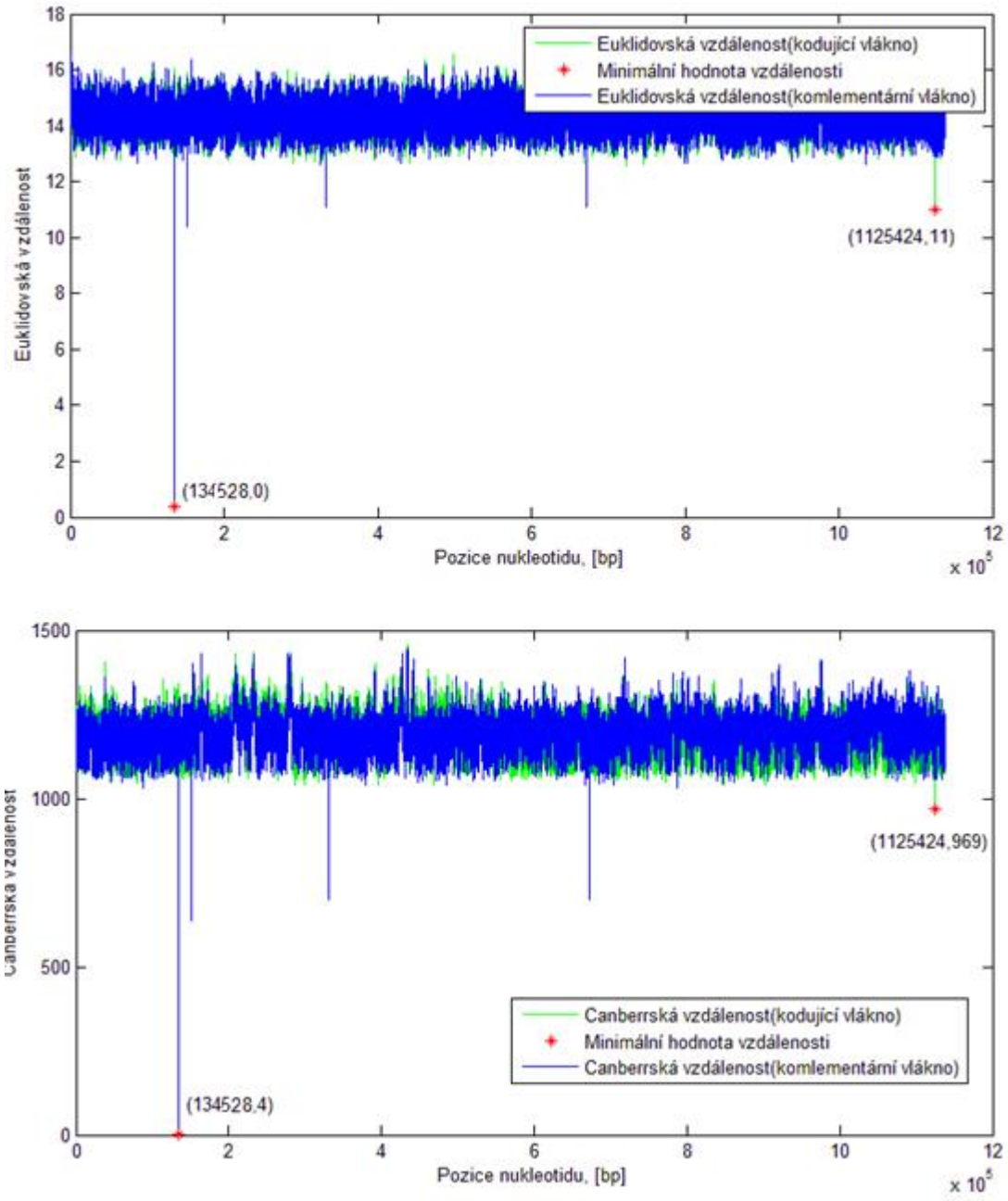
Obr. 5.9– Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 4. gen. Numerická reprezentace: denzitní vektory s oknem 3.



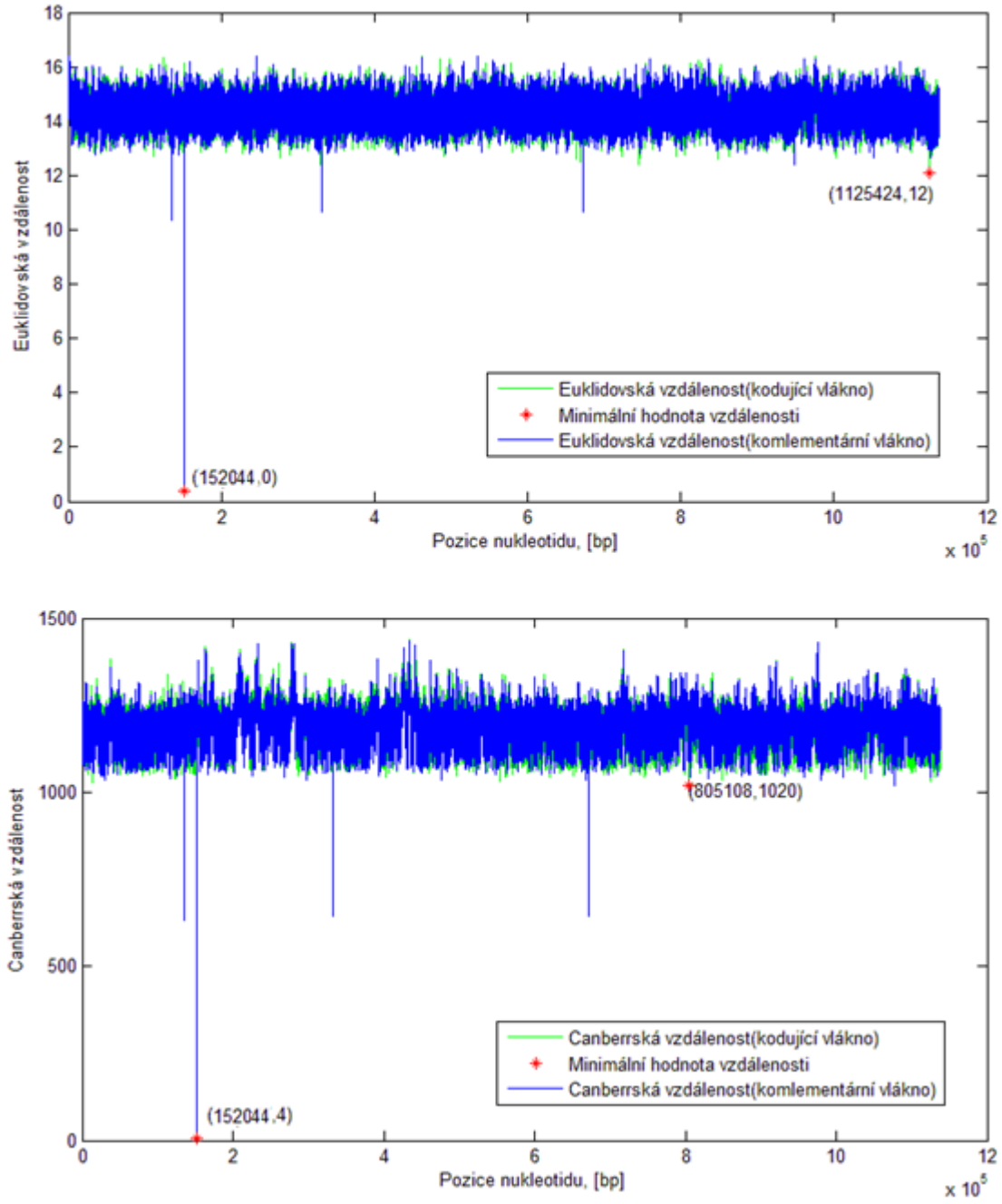
Obr. 5.10 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 5. gen. Numerická reprezentace: denzitní vektory s oknem 3.



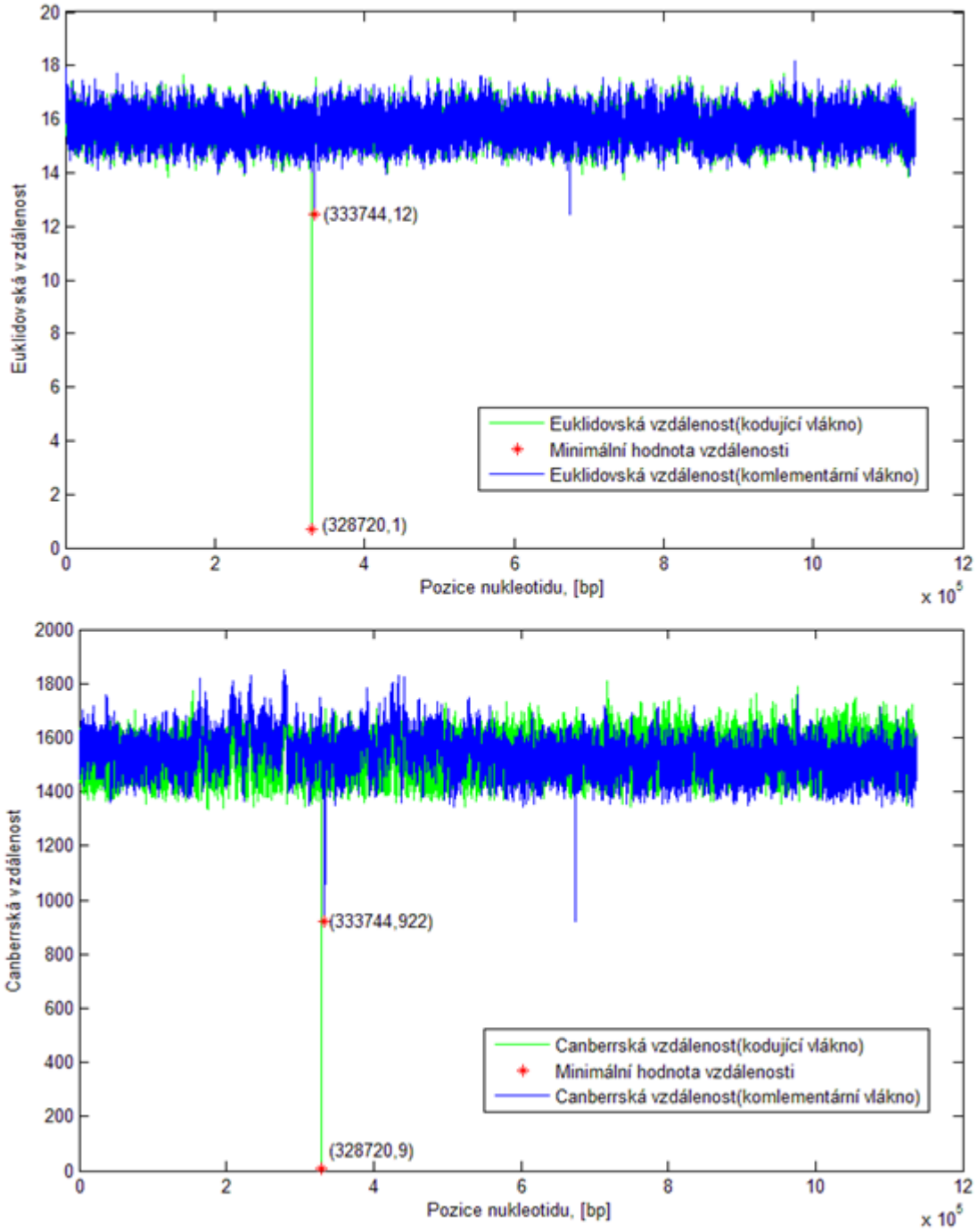
Obr. 5.11 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 5.



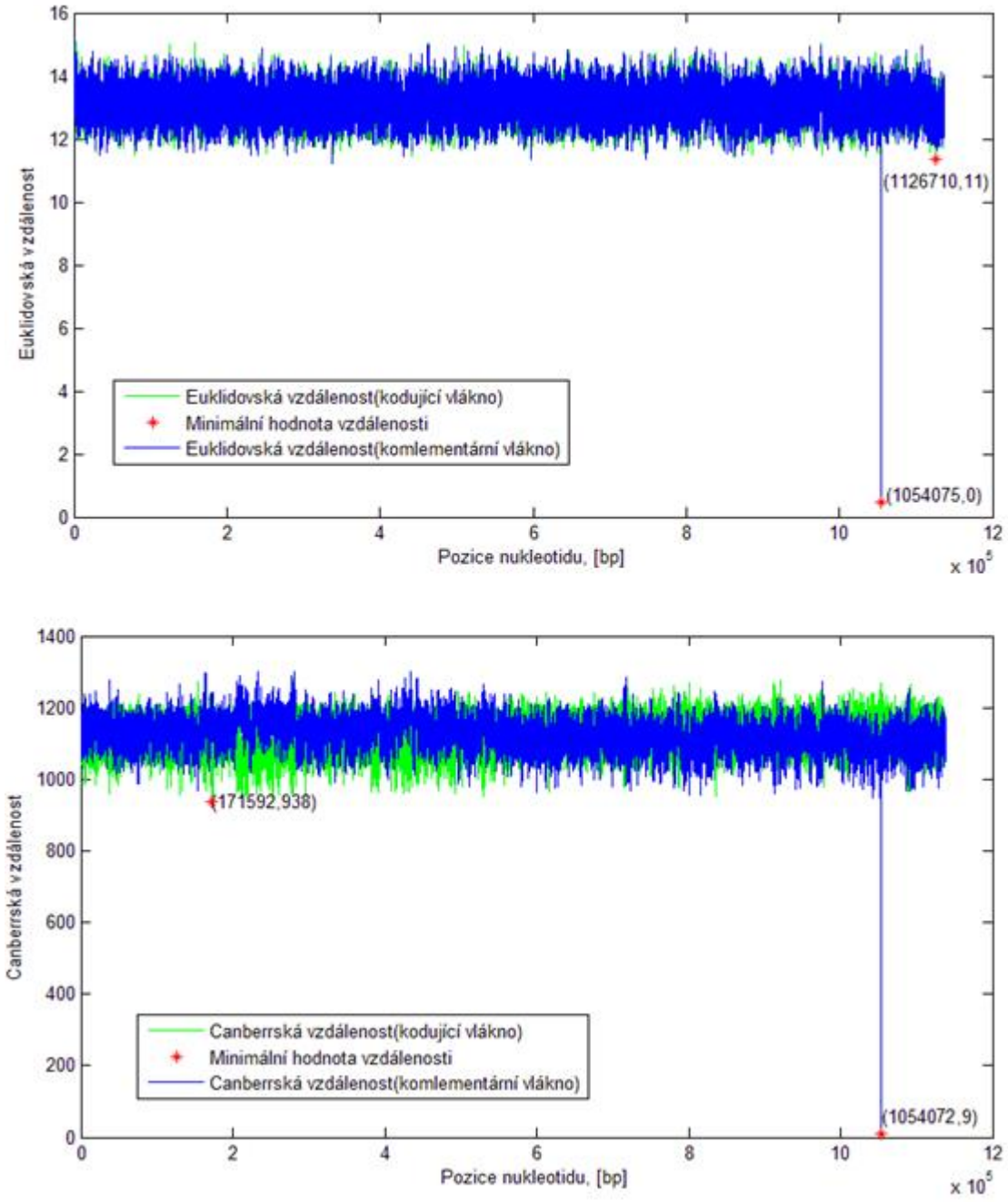
Obr. 5.12 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 2. gen. Numerická reprezentace: denzitní vektory s oknem 5.



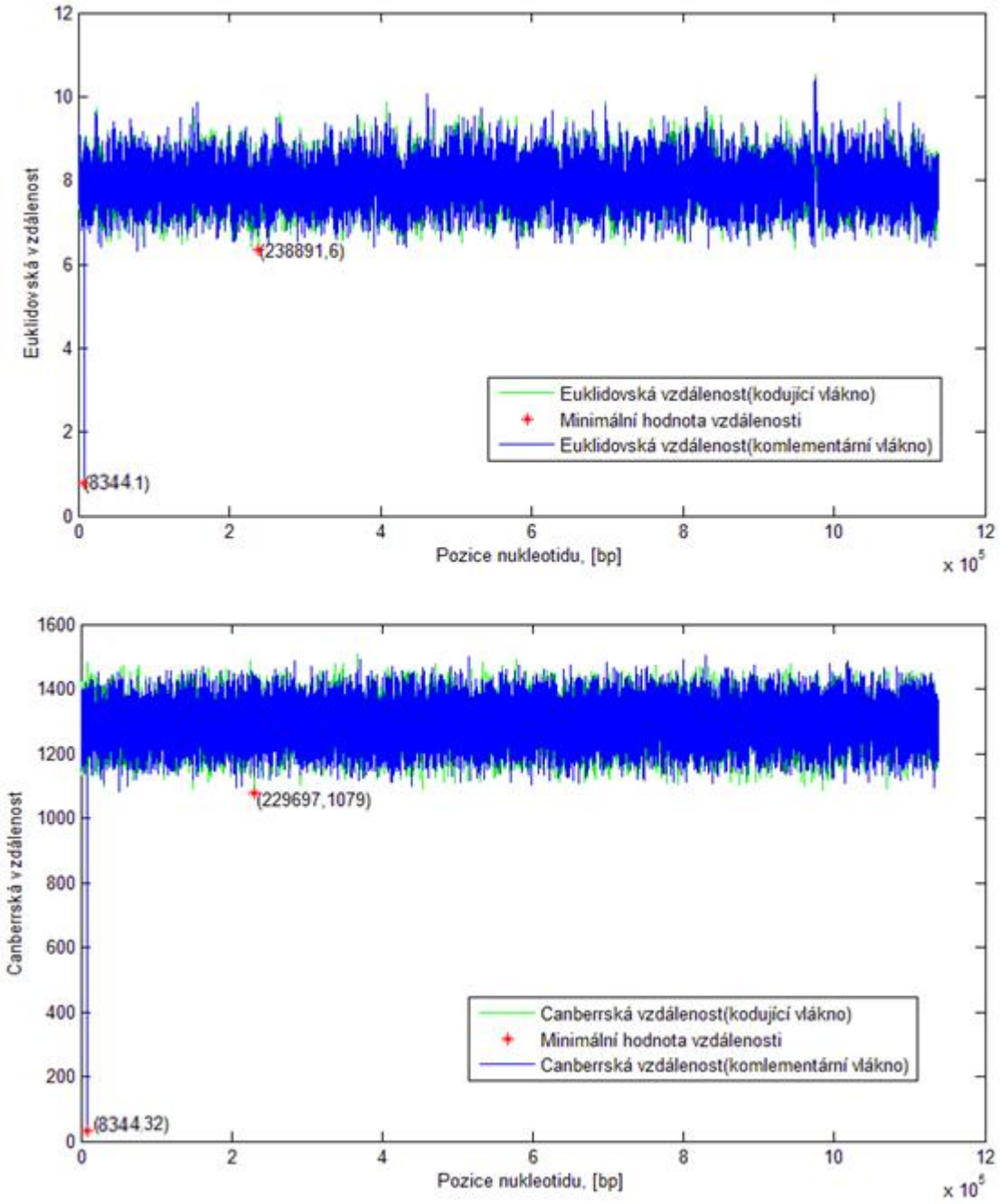
Obr. 5.13 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 3. gen. Numerická reprezentace: denzitní vektory s oknem 5.



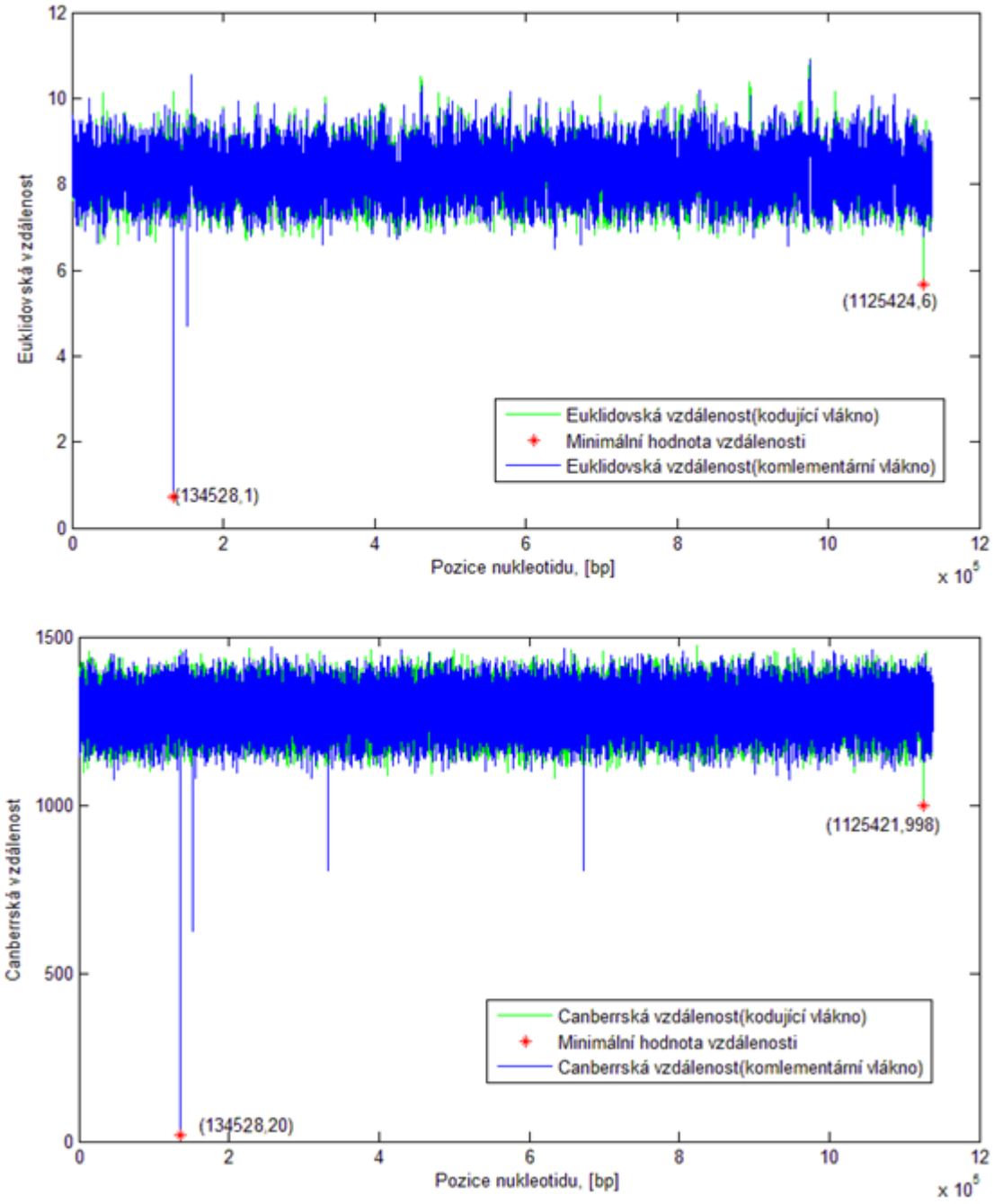
Obr. 5.14 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 4. gen. Numerická reprezentace: denzitní vektory s oknem 5.



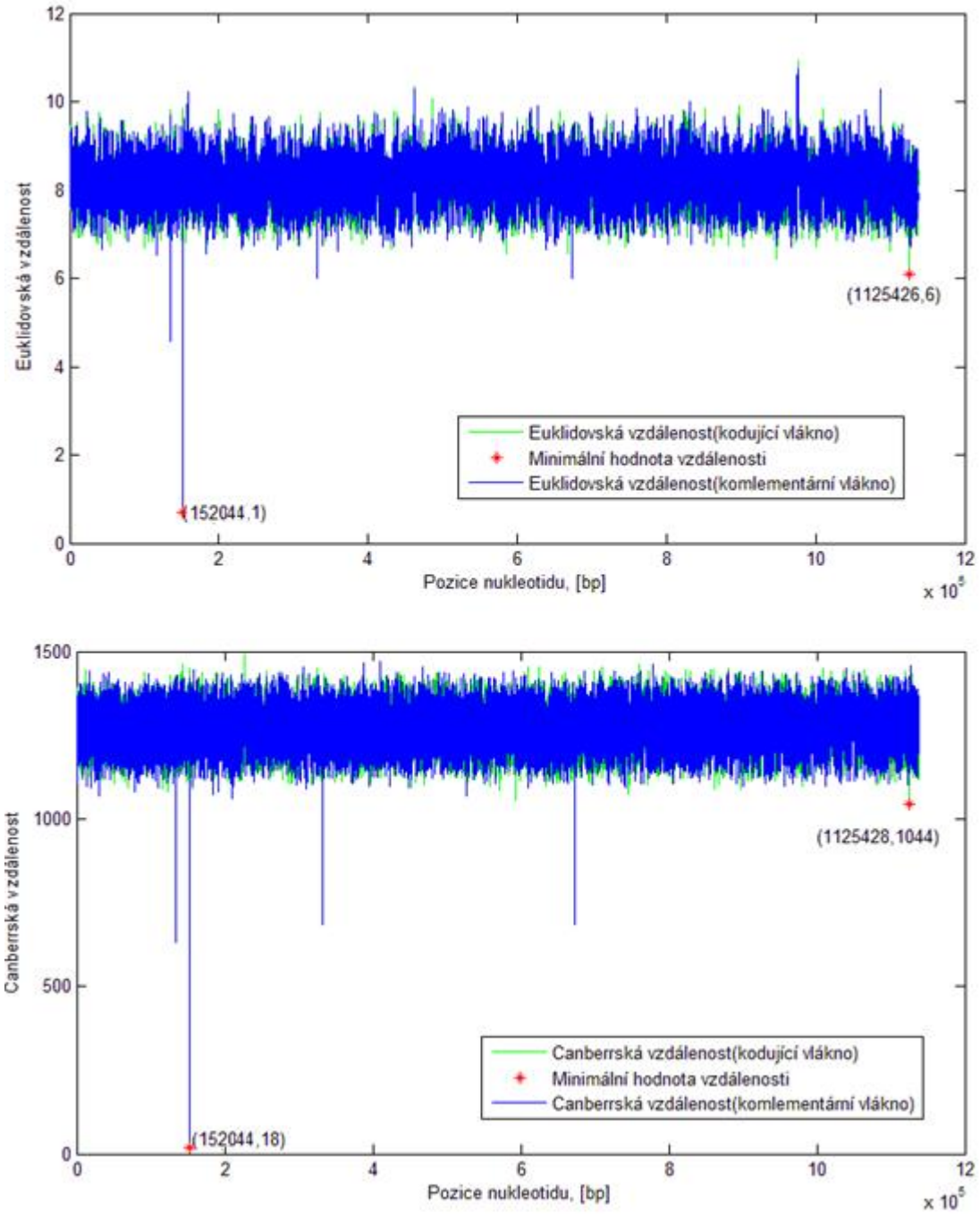
Obr. 5.15 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 5. gen. Numerická reprezentace: denzitní vektory s oknem 5.



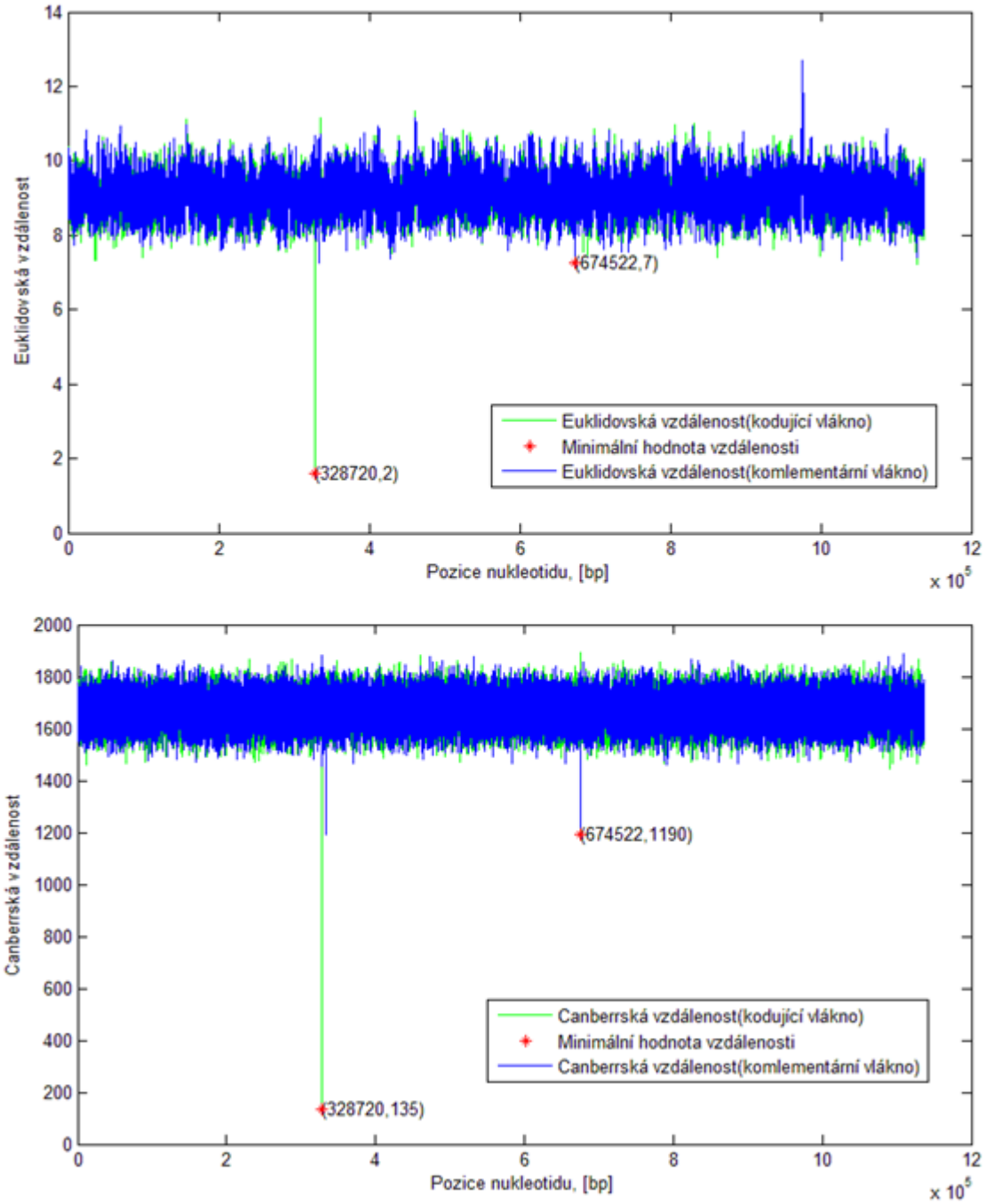
Obr. 5.16 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 17.



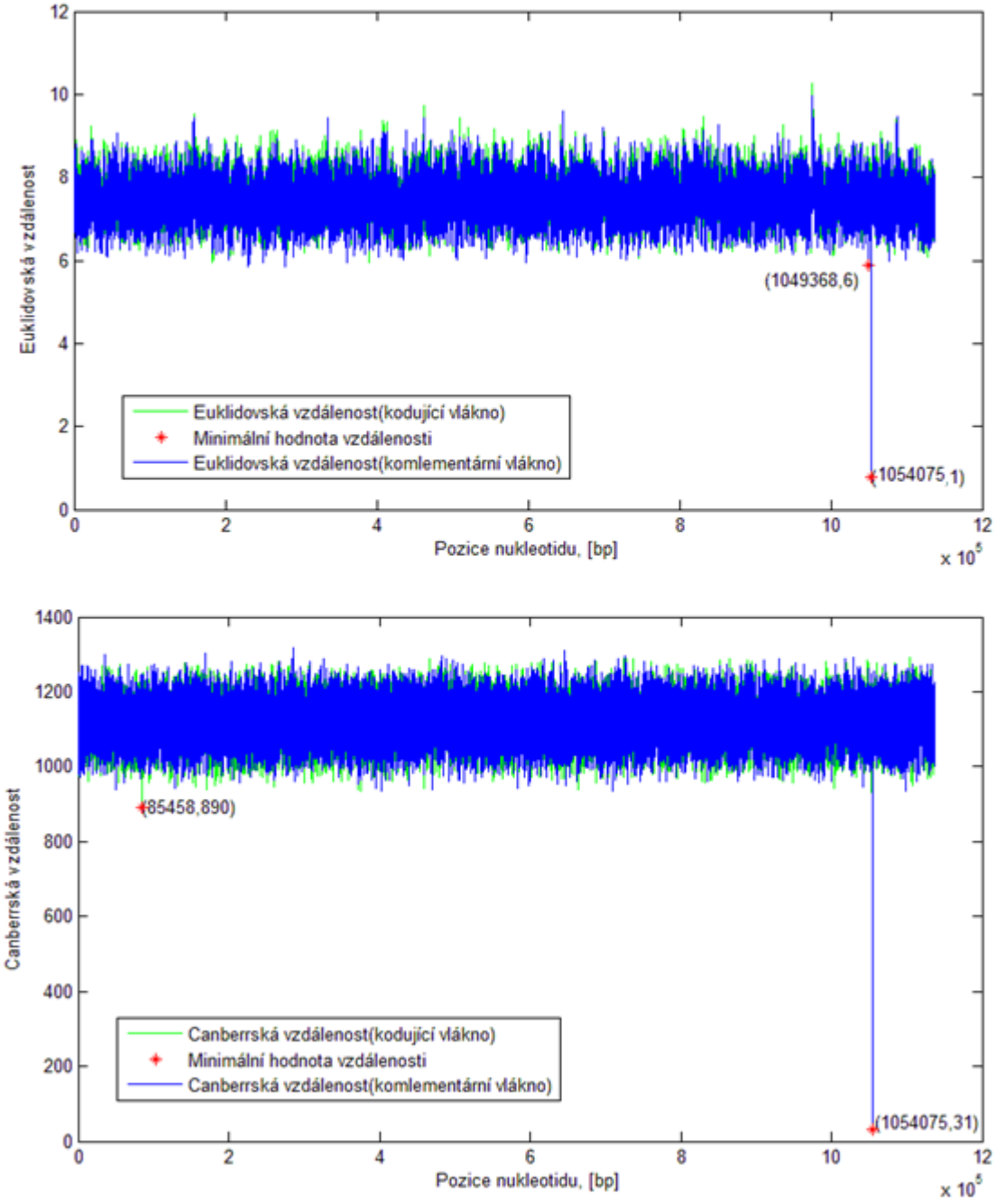
Obr. 5.17 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 2. gen. Numerická reprezentace: denzitní vektory s oknem 17.



Obr. 5.18 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 17.



Obr. 5.19 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 17.



Obr. 5.20 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: denzitní vektory s oknem 17.

Treponema pallidum subsp. pertenue str. CDC2

Podle v tab. 5.2 je vidět, že i pro tento poddruh organismu *treponema pallidum* výsledky vzdálenosti získané různými metodami jsou stejné. Což zvyšuje pravděpodobnost, že metoda detekovala homologní geny správně.

Grafy pro tento a další dva organismy nebudou ve práci ukázané, protože jsou velice podobné minulému organismu. Lze je nalézt v příloze spolu s kódem metody.

Rozbalená fáze v případě euklidovské vzdálenosti má rozsah hodnot od 0 do 97. V případě canberrské vzdálenosti rozsah je 0 až 1032. Nulová hodnota byla nalezená pro 5. gen, což znamená, že poddruh *CDC2* a *Samoa D* mají stejný gen kódující protein vnější membrány. Jiné organismy nulovou hodnotu vzdálenosti neměli.

Denzitní vektory vykazují stejnou tendenci jak i v předcházejícím případě: rozsah hodnot s rostoucí velikosti okna klesá (u euklidovské vzdálenosti). Pro okna 3, 5, 17 rozsahy hodnot pro euklidovskou vzdálenost jsou: 0,5 až 16, 0,4 až 12, 0,7 až 7 . Pro canberrskou naopak rozsah roste: 3 až 996, 4 až 1056, 18 až 1189.

Třeba také poznamenat, že minimální hodnoty vzdálenosti má to vlákno(kódující nebo komplementární), na kterém je původní gen.

Tabulka 5.2 – Výsledky pro *treponema pallidum subsp. pertenue str. CDC2*

Minimální euklidovská vzdálenost				**Gen**
RF	DV (3)	DV (5)	DV (17)	
224518	848965	864330	239237	**1.**
8344	8344	8344	8344	**1.(komp.)**
1125838	1125838	1125838	1125838	**2.**
134901	134901	134901	134901	**2.(komp.)**
663586	663586	1125838	1125840	**3.**
152346	152346	152346	152346	**3.(komp.)**
329056	329056	329056	329056	**4.**
334080	334080	334080	674976	**4.(komp.)**
544	885921	1127124	1049764	**5.**
1054486	1054486	1054486	1054486	**5.(komp.)**

Minimální canberrská vzdálenost				**Gen**
RF	DV(3)	DV(5)	DV(17)	
663481	839403	15166	230033	**1.**
8344	8344	8344	8344	**1.(komp.)**
1125838	1125838	1125838	1125835	**2.**
134901	134901	134901	134901	**2.(komp.)**
1125844	663586	805562	1125842	**3.**
152346	152346	152346	152346	**3.(komp.)**
329056	329056	329056	329056	**4.**
334080	334080	334080	674976	**4.(komp.)**
252105	172806	171926	85761	**5.**
1054486	1054486	1054486	1054486	**5.(komp.)**

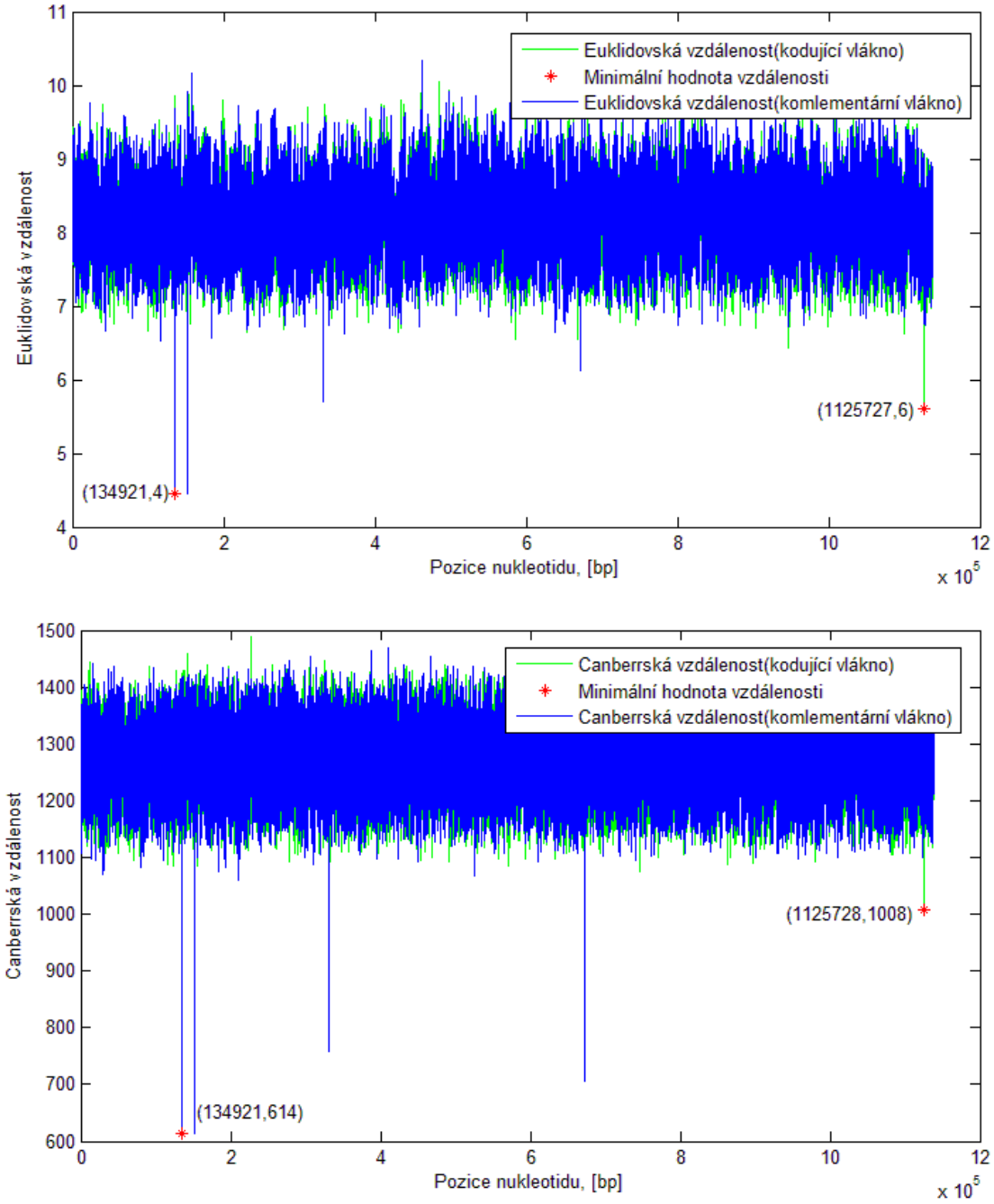
`

Treponema pallidum subsp. pallidum str. Nichols

Výsledky u tohoto organismu jsou taky stejné pro vše metody vyhledaní. Výjimku tvoří euklidovská vzdálenost denzitních vektorů s oknem 17. Důvod proč hodnota je jiná je patrná z Obr. 5.21.: dvě hodnoty vzdálenosti jsou k sobě natolik blízko, že nadměrné kmitaní denzitních vektoru s velkým oknem změnilo menší vzdálenost na větší. Proto by bylo vhodné používat pro méně podobné sekvence denzitní vektory s co nejmenším oknem. Na druhou stranu velikost má vliv na výpočetní náročnost. Metoda s malým oknem může být více účinná ale nebude možné ji využit na rozsáhlejších datech.

Pro euklidovskou vzdálenost rozsah rozbalené fáze je od 7 do 97. Pro canberrskou: 7 až 1029. Pro okna 3, 5, 17 denzitních vektorů rozsahy hodnot euklidovské vzdálenosti jsou: 1 až 16, 1 až 13, 1 až 6. U canberrské vzdálenosti rozsahy jsou následující: 4 až 997, 8 až 1056, 62 až 1082.

`



Obr. 5.21 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 3. gen. Numerická reprezentace: denzitní vektory s oknem 17.

Tabulka 5.3 – Výsledky pro treponema pallidum subsp. pallidum str. Nichols

Minimální euklidovská vzdálenost				**Gen**
RF	DV (3)	DV (5)	DV (17)	
224570	848527	863892	239279	**1.**
8344	8344	8344	8344	**1.(komp.)**
1125727	1125727	1125727	1125727	**2.**

134912	134912	134912	152314	**2.(komp.)**
1125727	1125727	1125727	1125727	**3.**
134918	134918	134918	134918	**3.(komp.)**
329144	329144	329144	329144	**4.**
333537	333537	333537	674540	**4.(komp.)**
544	885483	1127013	1049288	**5.**
1053998	1053998	1053998	1053998	**5.(komp.)**

Minimální canberrská vzdálenost				**Gen**
RF	DV (3)	DV (5)	DV (17)	
663031	838965	15164	230085	**1.**
8344	8344	8344	8344	**1.(komp.)**
1125727	1125727	1125727	1125727	**2.**
134912	134912	134912	134912	**2.(komp.)**
1125727	1125727	1125727	1125728	**3.**
134918	134918	134918	134918	**3.(komp.)**
329144	329144	329144	329144	**4.**
333537	333537	333537	674534	**4.(komp.)**
252147	172862	171982	85771	**5.**
1053998	1053998	1053998	1053998	**5.(komp.)**

Treponema pallidum subsp. pallidum str. Mexico A

Pro poslední poddruh organismu treponema pallidum také platí, že vše metody nalezli stejnou pozici homologního genu.

Rozsah vzdálenosti je podobný jako rozsah u poddruhu CDC 2, avšak tento druh nemá nulovou hodnotu (minimální hodnotu 5 má 4. gen).

Tabulka 5.4 – Výsledky pro treponema pallidum subsp. pallidum str. Mexico A

Minimální euklidovská vzdálenost				**Gen**
RF	DV (3)	DV (5)	DV (17)	
224613	848902	864267	239332	**1.**
8344	8344	8344	8344	**1.(komp.)**
1126132	1126132	1126132	1126132	**2.**
134913	134913	134913	134913	**2.(komp.)**
1126138	1126132	1126132	1126134	**3.**
152371	152371	152371	152371	**3.(komp.)**
329185	329185	329185	329185	**4.**
333578	674915	333578	674921	**4.(komp.)**
975964	885858	1127418	1049684	**5.**
1054403	1054403	1054403	1054403	**5.(komp.)**

`

Minimální canberrská vzdálenost				Gen
RF	DV (3)	DV (5)	DV (17)	
663410	839340	15163	230128	**1.**
8344	8344	8344	8344	**1.(komp.)**
1126132	1126132	1126132	1126132	**2.**
134913	134913	134913	134913	**2.(komp.)**
1126132	1126132	1126132	1126133	**3.**
152371	152371	152371	152371	**3.(komp.)**
329185	329185	329185	329185	**4.**
333578	333578	333578	674915	**4.(komp.)**
252200	172903	172023	85773	**5.**
1054403	1054403	1054403	1054403	**5.(komp.)**

Treponema paraluiscuniculi Cuniculi A

Treponema paraluiscuniculi Cuniculi A není poddruhem organismu treponema pallidum. Proto lze očekávat ještě před zahájením testovaní, že funkci některých hledaných proteinů můžou plnit proteiny z jiných skupin. Ve výsledku pak s použitím prahu nebude nalezen žádný homologní gen.

Výsledky v tab. 5.5 pro 1., 3. a 5. geny jsou stejné při použití všech metod. Na obr. 5.22 – 5.24 jsou ukázané vybrané způsoby vyhledání. Z obr. 5.23 a 5.24 je patrné, že signál je podobný předcházejícím u organizmů treponema pallidum, ale práh který byl navržen pro více příbuzné organismy není postačující pro 1. gen na obr. 5.22. Např. canberrská vzdálenost 426 přesahuje hodnotu 400. Otázka je, jestli počítat tento gen jako homologní nebo ne.

Odlišná situace nastává u 2. a 4. genů (viz Obr. 5.25 a 5.26). V případě 2. genu euklidovská vzdálenost 14 a canberrská 726 jsou příliš velké pro to, any tento gen byl označen jako homologní.

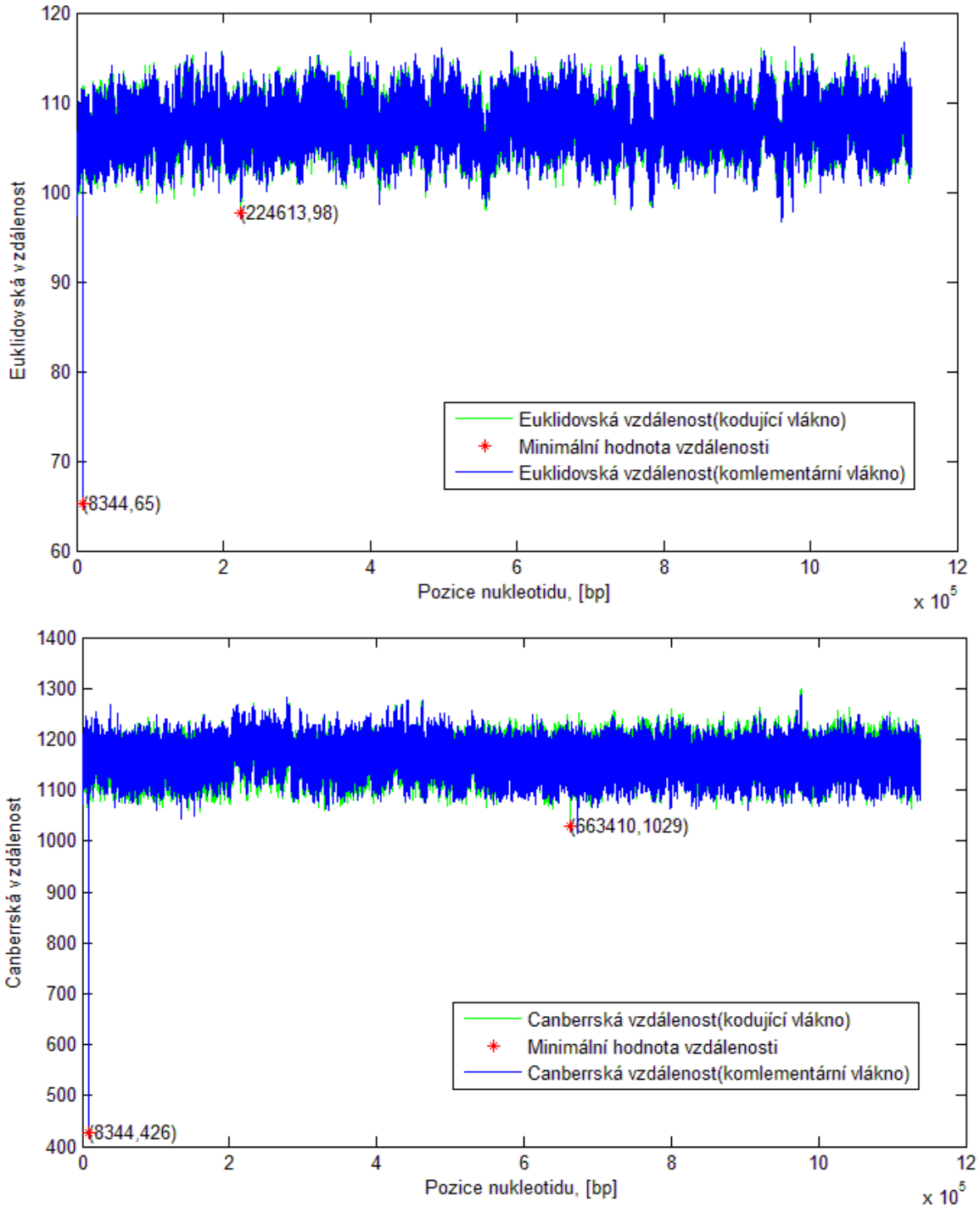
Tabulka 5.5 – Výsledky pro treponema paraluiscuniculi Cuniculi A

Minimální euklidovská vzdálenost				Gen
RF	DV (3)	DV (5)	DV (17)	
436	841781	857147	942180	**1.**
8345	8345	8345	8345	**1.(komp.)**
1119475	1119475	1119475	1119477	**2.**
134846	134841	134841	134841	**2.(komp.)**
1119475	659025	738827	940446	**3.**
134846	134846	134846	150800	**3.(komp.)**
326502	326519	326519	1121525	**4.**
667819	329202	329202	329202	**4.(komp.)**
568	1120766	1120767	1042991	**5.**
1047696	1047696	1047696	1047696	**5.(komp.)**

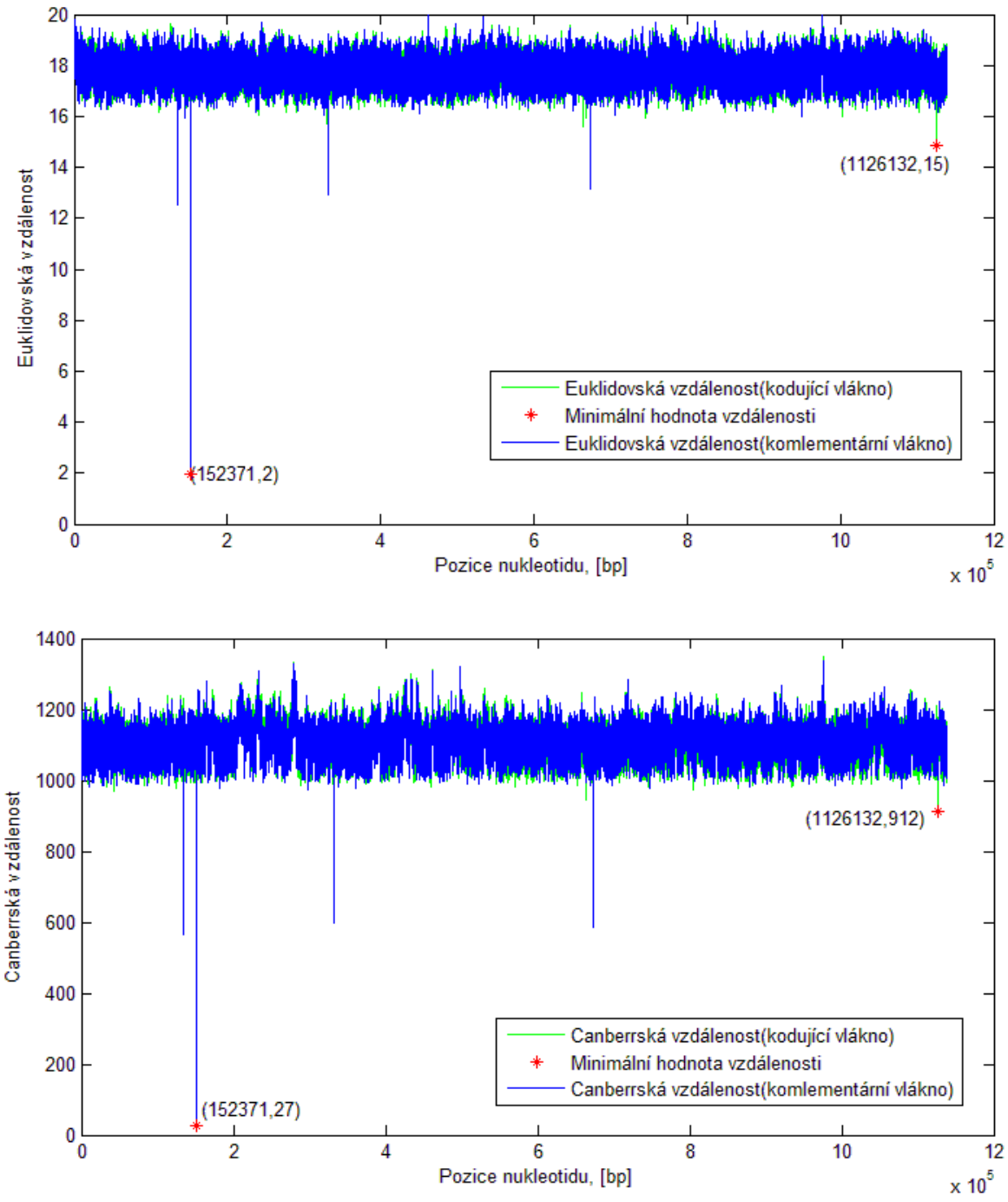
`

Minimální canberrská vzdálenost				Gen
RF	DV (3)	DV (5)	DV(17)	
658920	8612	153453	227508	**1.**
8345	8345	8345	8345	**1.(komp.)**
1119475	1119475	1119475	1119478	**2.**
134846	134841	134841	150795	**2.(komp.)**
1119475	659025	798386	589606	**3.**
134846	134846	134846	134846	**3.(komp.)**
326502	326519	326519	326502	**4.**
667819	329202	667802	329202	**4.(komp.)**
249571	170272	169392	85861	**5.**
1047696	1047696	1047696	1047696	**5.(komp.)**

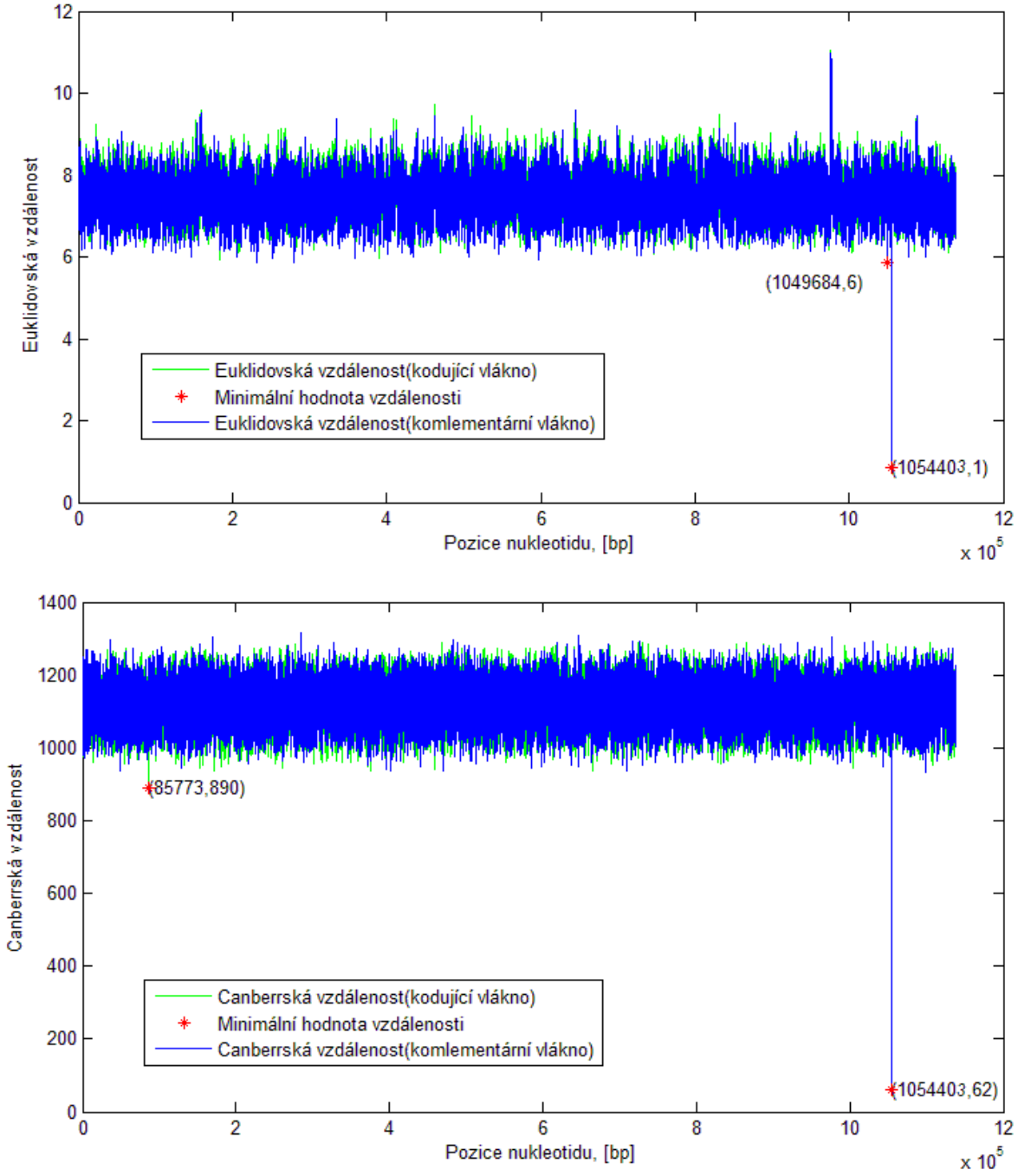
`



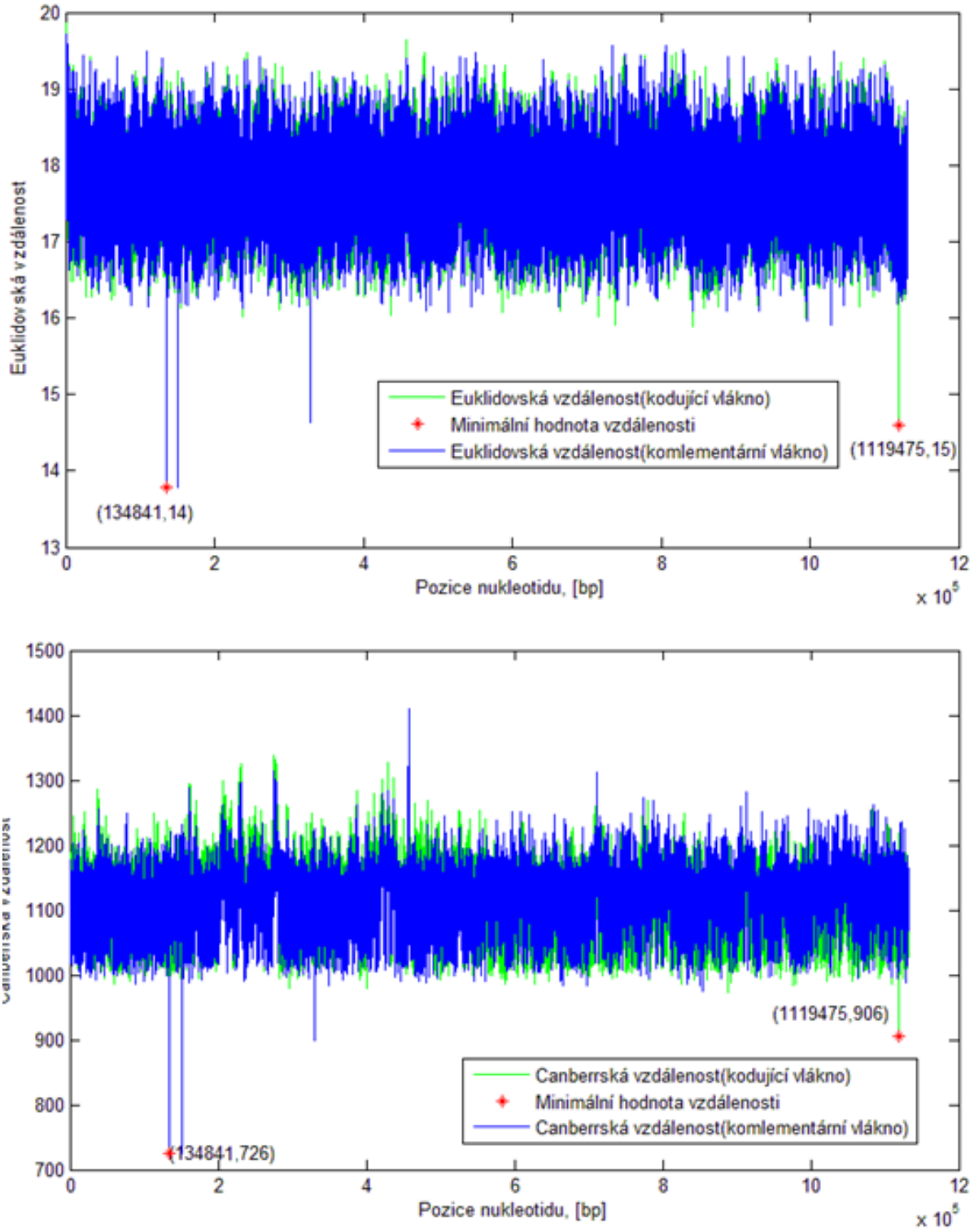
Obr. 5.22 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 1. gen. Numerická reprezentace: rozbalená fáze.



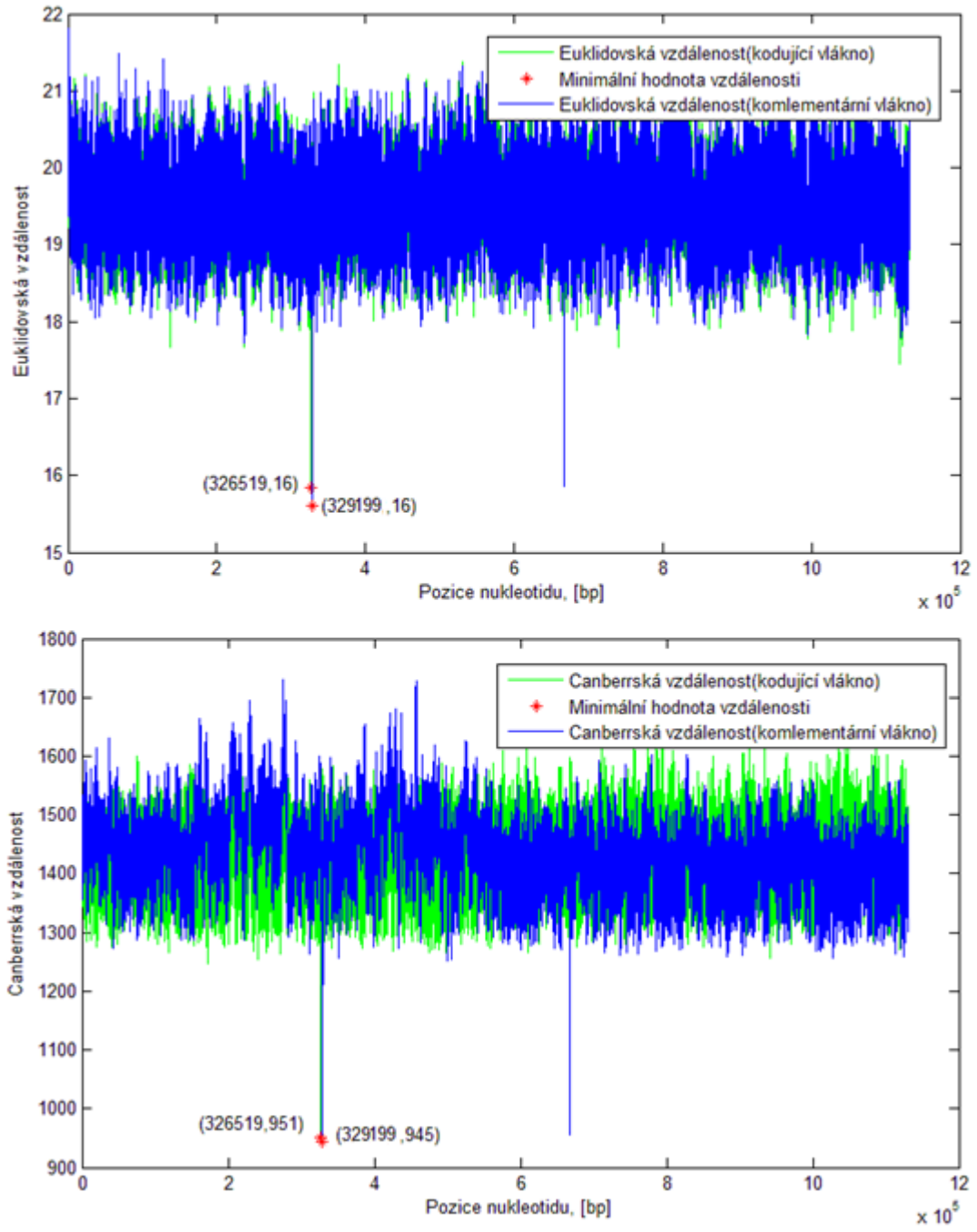
Obr. 5.23 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 3. gen. Numerická reprezentace: denzitní vektory s oknem 3.



Obr. 5.24 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 5. gen. Numerická reprezentace: denzitní vektory s oknem 17.



Obr. 5.25 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 2. gen. Numerická reprezentace: denzitní vektory s oknem 3.



Obr. 5.26 – Euklidovská (nahoře) a canberrská vzdálenost (dole) pro 5. gen. Numerická reprezentace: denzitní vektory s oknem 3.

ZÁVĚR

Homologní geny DNA sekvence můžou být vyhledávané v databázích za účelem zjištění její funkce, protože geny mající společného předka mohou mít podobné vlastnosti.

Prvním krokem je převod nukleové posloupnosti DNA nebo RNA na číslicový signál. Numerická mapa, kterou by se dálo použit pro další analýzu, musí splňovat několik požadavků: odrážet biochemické vlastnosti nukleotidů, nezavádět chybnou informaci, být výpočetně nenáročnou. Jako vhodné byli vybrané metody s použitím fázové charakteristiky komplexních čísel a reprezentace denzitními vektory. Obě metody se ukázali jako vhodné při vyhledání CDS úseku v celém genomu.

Dalším krokem je zvolení vhodné metody zpracování signálu. Korelace a většina metrik nebyli použité, protože měli charakteristiky, které by nezbytně vedli k chybám. Nejvhodnější metody (euklidovská a canberrská vzdálenosti) byli použité pro vytvoření programu. Euklidovská vzdálenost má menší rozsah a výsledný signál je hladší, než u canberrské vzdálenosti.

Podle výsledků je vidět, že nejvhodnější metodou byla canberrská vzdálenost při použití rozbalené fáze nebo denzitních vektoru s malým oknem(délka 3 nukleotidy). Canberrská vzdálenost má víceméně staly rozsah hodnot pro různé numerické reprezentaci, což dovoluje nastavit práh hodnot pro vše metody stejně nebo podobně. Euklidovskou vzdálenosti také lze dostat správnou polohu homologních genů.

Obě metody jsou velmi citlivé na natavení prahu. Práh třeba volit v závislosti na tom, zda se požaduje co nejvíce homologních genů nebo naopak jeden, u kterého tato podobnost je výraznější.

`

LITERATURA

[1] VOSS, Richard F. Evolution of long-range fractal correlations and 1/ f noise in DNA base sequences. Physical Review Letters [online]. 1992, 68(25), 3805-3808 [cit. 2017-01-29]. DOI: 10.1103/PhysRevLett.68.3805. ISSN 0031-9007. Dostupné z: http://link.aps.org/doi/10.1103/PhysRevLett.68.3805

[2] SILVERMAN, B.D. a R. LINSKER. A measure of DNA periodicity. Journal of Theoretical Biology [online]. 1986, 118(3), 295-300 . DOI: 10.1016/S0022-5193(86)80060-1. ISSN 00225193. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0022519386800601

[3] CRISTEA, Paul D., Manfred D. KESSLER a Gerhard J. MUELLER. [online]. In: . s. 77-84 [cit. 2017-01-29]. DOI: 10.1117/12.491244. Dostupné z: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=872138

[4] ZHAO, Jing, Xiu Wen YANG, Jian Ping LI a Yuan Yan TANG. DNA Sequences Classification Based on Wavelet Packet Analysis [online]. s. 424 [cit. 2017-01-29]. DOI: 10.1007/3-540-45333-4_53. Dostupné z: http://link.springer.com/10.1007/3-540-45333-4_53

[5] ANASTASSIOU, D. Genomic signal processing. IEEE Signal Processing Magazine [online]. 18(4), 8-20 . DOI: 10.1109/79.939833. ISSN 10535888. Dostupné z: http://ieeexplore.ieee.org/document/939833/

[6] MADĚRÁNKOVÁ, D. Analýza nukleotidových denzit jako metoda pro identifikaci organismů. In Nové směry v biomedicínském inženýrství. Brno: 2013. s. 8-13. ISBN: 978-80-214-4814-8.

[7] Yau SS-T, Wang J, Niknejad A, Lu C, Jin N, Ho Y-K. DNA sequence representation without degeneracy. Nucleic Acids Research. 2003;31(12):3078-3080.

[8] AKHTAR, Mahmood, Julien EPPS a Eliathamby AMBIKAIRAJAH. On DNA Numerical Representations for Period-3 Based Exon Prediction. In: 2007 IEEE International Workshop on Genomic Signal Processing and Statistics [online]. IEEE, 2007, s. 1-4 [cit. 2017-01-29]. DOI: 10.1109/GENSIPS.2007.4365821. ISBN 978-1-4244-0998-3. Dostupné z: http://ieeexplore.ieee.org/document/4365821/

[9] ABO-ZAHHAD, Mohammed, Sabah M. AHMED a Shimaa A. ABD-ELRAHMAN. Integrated Model of DNA Sequence Numerical Representation and Artificial Neural Network for Human Donor and Acceptor Sites Prediction. International Journal of Information Technology and Computer Science [online]. 2014, 6(8), 51-57 [cit. 2017-01-29]. DOI: 10.5815/ijitcs.2014.08.07. ISSN 20749007. Dostupné z: http://www.mecs-press.org/ijitcs/ijitcs-v6-n8/v6n8-7.html

[10] Todd Holden, R. Subramaniam, R. Sullivan, E. Cheng, C. Sneider, G. Tremberger, Jr. A. Flamholz, D. H. Leiberman, and T. D. Cheung, "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes," in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE), vol. 6694, August 2007, pp. 669417-1 to 669417-10.

[11] BERGER, John A, Sanjit K MITRA, Marco CARLI a Alessandro NERI. Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute [online]. 2004, 341(1-2), 37-53 [cit. 2017-01-29]. DOI:

`

10.1016/j.jfranklin.2003.12.002. ISSN 00160032. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0016003203000917

[12] ZHANG, Ren a Chun-Ting ZHANG. Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences. Journal of Biomolecular Structure and Dynamics [online]. 1994, 11(4), 767-782 [cit. 2017-01-29]. DOI: 10.1080/07391102.1994.10508031. ISSN 0739-1102. Dostupné z: http://www.tandfonline.com/doi/abs/10.1080/07391102.1994.10508031

[13] ALTSCHUL, Stephen F., Warren GISH, Webb MILLER, Eugene W. MYERS a David J. LIPMAN. Basic local alignment search tool. Journal of Molecular Biology [online]. 1990, 215(3), 403-410 [cit. 2017-01-29]. DOI: 10.1016/S0022-2836(05)80360-2. ISSN 00222836. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602

[14] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST a PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 : 3389 až 3402.

[15] [online] Dostupné z: http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

[16] SIEVERS, F., A. WILM, D. DINEEN, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology [online]. 2011, 7(1), 539-539 [cit. 2017-01-29]. DOI: 10.1038/msb.2011.75. ISSN 1744-4292. Dostupné z: http://msb.embopress.org/cgi/doi/10.1038/msb.2011.75

[17] PEARSON, William R. An Introduction to Sequence Similarity ("Homology") Searching. Current Protocols in Bioinformatics [online]. Hoboken, NJ, USA: John Wiley & Sons, 2002 [cit. 2017-01-29]. DOI: 10.1002/0471250953.bi0301s42. ISBN 0471250953. Dostupné z: http://doi.wiley.com/10.1002/0471250953.bi0301s42

[18] JAN, Jiří. Číslicová filtrace, analýza a restaurace signálů. 2. upr. a rozš. vyd. Brno: VUTIUM, 2002. ISBN 80-214-1558-4.

[19] RAJARAMAN, Anand. a Jeffrey D. ULLMAN. *Mining of massive datasets*. Cambridge: Cambridge University Press, 2012. ISBN 1107015359.

[20] HARUŠTIAKOVÁ, Danka. *Vícerozměrné statistické metody v biologii*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-791-8.

[21] HOLČÍK, Jiří. *Analýza a klasifikace dat*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-793-2.

`